

**EFFECT SIZE MATTERS: EMPIRICAL INVESTIGATIONS TO HELP
RESEARCHERS MAKE INFORMED DECISIONS ON COMMONLY USED
STATISTICAL TECHNIQUES**

A Dissertation

by

SUSANA TRONCOSO SKIDMORE

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

December 2009

Major Subject: Educational Psychology

**EFFECT SIZE MATTERS: EMPIRICAL INVESTIGATIONS TO HELP
RESEARCHERS MAKE INFORMED DECISIONS ON COMMONLY USED
STATISTICAL TECHNIQUES**

A Dissertation

by

SUSANA TRONCOSO SKIDMORE

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Bruce Thompson
Committee Members,	Robert M. Capraro
	Oi-man Kwok
	Victor L. Willson
Head of Department,	Victor L. Willson

December 2009

Major Subject: Educational Psychology

ABSTRACT

Effect Size Matters: Empirical Investigations to Help Researchers Make Informed
Decisions on Commonly Used Statistical Techniques. (December 2009)

Susana Troncoso Skidmore, B.A., M.Ed., Texas A&M University

Chair of Advisory Committee: Dr. Bruce Thompson

The present journal article formatted dissertation assessed the characteristics of effect sizes of commonly used statistical techniques. In the first study, the author examined the *American Educational Research Journal (AERJ)* and select *American Psychological Association (APA)* and *American Counseling Association (ACA)* journals to provide an historical account and synthesis of which statistical techniques were most prevalent in the fields of education and psychology. These reviews represented a total of 17,698 techniques recorded from 12,012 articles. Findings point to a general decrease in the use of the *t*-test and ANOVA/ANCOVA and a general increase in the use of regression and factor/cluster analysis.

In the second study, the author compared the efficacy of one Pearson r^2 and seven multiple R^2 correction formulas for the Pearson r^2 . The author computed adjustment bias and precision under 108 conditions (6 population ρ^2 values, 3 shape conditions and 6 sample size conditions). The Pratt and the Olkin-Pratt Extended formulas more consistently provided unbiased estimates across sample sizes, ρ^2 values and the shape conditions investigated.

In the third study, the author evaluated the robustness of estimates of practical significance (η^2 , ε^2 and ω^2) in one-way between subjects univariate ANOVA. There were 360 simulation conditions (5 population Cohen's d values, 4 group proportion ratios, 3 shape conditions, 3 variance conditions, and 2 total sample size conditions) for each of three group configurations (2, 3 and 4 groups). Three indices of practical significance (η^2 , ε^2 , ω^2) and two indices of statistical significance (Type I error and power) were computed for each of the 5,400, 000 (5,000 replications x 360 simulation conditions x 3 group configurations). Simulation findings for η^2 under heterogeneous variance conditions indicated that for the $k=2$ and $k=3$ condition Cohen's d values up to 0.2 (up to 0.5 for $k=4$) tend to produce overestimated population η^2 values. Under heterogeneous variance conditions for ε^2 and ω^2 at Cohen's $d = 0.0$ and 0.2 , the negative variance pairing overestimated and the positive variance pairing underestimated the parameter η^2 but at Cohen's $d \geq 0.5$, both the positive and negative variance conditions resulted in underestimated parameter η^2 values.

DEDICATION

This document is dedicated to Austin and Dallas, my beautiful and precious gifts from God. “Be joyful always; pray continually; give thanks in all circumstances, for this is God's will for you in Christ Jesus” (1 Thessalonians 5: 16-18, New International Version).

ACKNOWLEDGEMENTS

The most important acknowledgement is always Jesus Christ, my Lord and Savior, for His mercy, His grace and His countless blessings. “Praise be to the Lord, to God our Savior, who daily bears our burdens” (Psalm 68:19). “ But thanks be to God, who always leads us in triumphal procession in Christ and through us spreads everywhere the fragrance of the knowledge of him” (2 Corinthians 2:14).

There are many people that helped me throughout this dissertation process, but undoubtedly the person that provided the most professional guidance was Dr. Bruce Thompson, my committee chair. Dr. Thompson, I have the deepest respect and admiration for you. You work harder than any professor I have ever known to help your students succeed. You introduced me to a way of thinking and questioning beyond statistics that has captivated my attention and made me want to learn even more. Most importantly, you pushed me out of my comfort zone, to take a leap of faith and accomplish what I thought I could not possibly accomplish. Since I became your student, I have had the opportunity to teach, publish, serve on committees, travel all over the country (even out of the country!), be recognized with awards, etc. Undoubtedly, there is a strong relationship between the occurrence of these events and becoming your student ($\eta^2 = 77\%$, $p < 0.07$)! There are so many things to thank you for, but most of all I am thankful that you lead by example. I have asked for your advice on many matters and your responses have always been that of a man of character and integrity. Whether directly through the words you speak in your classes, at workshops/training sessions, in

your office or indirectly through your countless written works, I never tire of learning from you. You are a brilliant writer, a disciplined researcher and a captivating teacher - in short - you are a true blessing in my life.

I would also like to thank my committee members, Dr. Victor L. Willson, Dr. Robert M. Capraro, and Dr. Oi-Man Kwok, for their guidance and support throughout the course of this research. I had the opportunity to have several classes with Dr. Willson and Dr. Kwok. I value each of your areas of expertise and hope that I will be able to make use of the advanced training I received from you both. Dr. Capraro and Dr. Kwok, I also want to thank you for the opportunity to work under you. Both work experiences were invaluable to my training and I very much appreciate your patience and support.

Although not committee members, I would also like to acknowledge Dr. Jim McNamara and Dr. Bob Hall. I will not forget the many lessons you taught me Dr. McNamara. Dr. Hall, thank you for contributing greatly to my 'stat library'!

I also want to extend my heartfelt gratitude to Dr. Becky Petitt and Dave McIntosh whom I had the honor of working with for a while. I treasure the laughter and good times we shared and fully appreciate the support, encouragement and advice you both graciously gave to me.

A most special thank you goes to Mark, my Taiwanese brother. I am so glad we were able to travel this dissertation road together. I have learned and continue to learn so much from you. I am going to be so sad for me (and happy for you) when you go to Taiwan!

A huge thank you also goes to my fellow grad student friends (Caroline, Christine, Ebrar, Eunju, Eun Sook, Helene, Jerry, June, Karen, Keisha, Laura, Lisa, Malena, Minjung, Myunghee, Robert, Ross, Russ, Toni, Yan, Yolanda and Yi-Chun). I enjoyed all the times we shared together - lunches, study times, class times, just hanging out at the EREL, etc. All I can say is - it's potluck time!!

I also want to tell Ra'sheedah Richardson how much she means to me. Our talks about faith, family and life were such a blessing to me. May God continue to bless and pour into your life so that you can continue to share your blessings with others and give Him all the glory.

I also need to thank my long time friend, Mel. Thank you for always being there for me girl. Love and hugs always.

It would have been very difficult to have gone through graduate school without the help and support of my mother and father in law. Thank you for always making yourselves available to help us.

To my brother, I love you - my favorite brother! I want to thank you and Michelle for always being available whenever and for whatever I needed you for.

To my momma and daddy, *los mas mejores padres en todo el universo*, words cannot express how much you both mean to me. Our daily talks, your prayers and words of encouragement have always been and continue to be my sustenance through the ups and downs of this life. I love you so much and hope that you know that I recognize that all my successes in life are because you have always kept me and Coni in your constant prayers.

Last but certainly not least, a lifetime debt of gratitude is owed to my husband, Vance. You are my best friend and my life partner. Thank you for your patience, love and understanding. I know it was difficult. I know you had to shoulder much more than you ever bargained for when we decided to do this grad school thing - but guess what - we did it baby! I love you!

TABLE OF CONTENTS

	Page
ABSTRACT	iii
DEDICATION.....	v
ACKNOWLEDGEMENTS	vi
TABLE OF CONTENTS	x
LIST OF FIGURES	xii
LIST OF TABLES	xiii
INTRODUCTION	1
Effect Sizes	4
Organization of Document	7
REVIEW OF REVIEWS OF STATISTICAL TECHNIQUES USED IN EDUCATION AND PSYCHOLOGY	9
Purpose	11
Methods	12
Results	17
Utility of Research Findings	31
Recommendations for Future Reviews of Statistical Techniques	33
CHOOSING THE BEST CORRECTION FORMULA FOR PEARSON r^2	35
Introduction.....	35
Methods	44
Results	46
Analysis	53
Conclusions and Recommendations.....	56

	Page
THE ROBUSTNESS OF ESTIMATES OF PRACTICAL SIGNIFICANCE IN ANOVA	59
Moving Beyond Statistical Significance Testing.....	62
Methods	66
Simulation Baseline Check.....	71
Analyses.....	74
Results	76
Discussion.....	89
SUMMARY AND CONCLUSIONS	92
Study One	92
Study Two.....	93
Study Three.....	94
REFERENCES	97
VITA.....	111

LIST OF FIGURES

	Page
Figure 1 Statistical Technique Reporting Trends in AERJ from 1969 to 1997 as a Percentage of Statistical Technique per Article.....	25
Figure 2 Statistical Technique Reporting Trends in Psychology Journals From 1948 to 2001 as a Percentage of Statistical Technique per Article.....	29
Figure 3 Statistical Technique Distributions in Psychology Journals as a Percentage of Statistical Technique per Article	32
Figure 4 Excel Sheet with Correction Formulas	42
Figure 5 Mean Biases of the Seven R^2 Correction Formulas Using the Simulation Design Features. Wherry Is Included So As to Be Able to Compare the Uncorrected r^2 Estimate	51
Figure 6 Empirical Estimates of Eta Squared and Omega Squared Under Variance Conditions by Group Size Configuration.....	85

LIST OF TABLES

	Page
Table 1 Patterns of Statistical Technique Coding Across Articles	15
Table 2 Introduction of Statistical Term by Year and Author	18
Table 3 Percentage of Techniques by Articles per Year in AERJ, 1969 – 1997	19
Table 4 How Studies Varied in Their Review of Articles	21
Table 5 Percentage of Statistical Techniques by Articles per Year in Psychology Journals, 1948 – 2001	27
Table 6 95% Confidence Intervals for Mean Percentage of Statistical Usage per Article from 1990- 1997	30
Table 7 Six R^2 Correction Formulas and a Pearson r^2 Correction Formula with Excel Formula Syntax	41
Table 8 SPSS Syntax	43
Table 9 Adjusted r^2 Values for Selected Sample Sizes and Uncorrected r^2 Values	45
Table 10 Mean Bias (Estimate - Parameter ρ^2) Across 5,000 Samples Drawn in Each of 108 Simulation Conditions	47
Table 11 Statistics for Bias of the 540,000 Values for Each of the Eight Estimates Across the 108 (6 X 3 X 6) Simulation Conditions ($n = 5,000$ /Condition)	52
Table 12 Statistics for Precision (SD) for Each of the Eight Estimates Within the 108 (6 X 3 X 6) Simulation Conditions ($n = 5,000$ /Condition)...	54

	Page
Table 13 Percentage of Mean Unbiased Estimates Returned Within the 108 (6 X 3 X 6) Simulation Conditions ($n = 5,000/\text{Condition}$)	55
Table 14 Empirical Type I Error Rates	72
Table 15 Empirical Power Estimates (Normal Distribution and Equal Variances).....	73
Table 16 Impact of Heterogeneity of Variance on Type I Error Rates	75
Table 17 Estimated Parameter Bias in Three ANOVA Effect Sizes for Two Groups.....	77
Table 18 Estimated Bias in Three ANOVA Effect Sizes for Three Groups	78
Table 19 Estimated Bias in Three ANOVA Effect Sizes for Four Groups	79
Table 20 Estimated Absolute Bias in Three ANOVA Effect Sizes for Two Groups.....	81
Table 21 Estimated Absolute Bias in Three ANOVA Effect Sizes for Three Groups.....	82
Table 22 Estimated Absolute Bias in Three ANOVA Effect Sizes for Four Groups.....	84
Table 23 Two Group Precision	87
Table 24 Three Group Precision	88
Table 25 Four Group Precision.....	89

INTRODUCTION

Researchers have surveyed published literature for research design practices (Bliss, Skinner, Hautau, & Carroll, 2008; Robinson, Levin, Thomas, Pituch, & Vaughn, 2007; Varnell, Murray, Janega, & Blitstein, 2004), measurement practices (Capraro, Capraro, & Henson, 2001; Vacha-Haase, Kogan, & Thompson, 2000) and statistical techniques and practices (Bowen, Rollins, Baggett, & Miller, 1990; Thompson, 1997; Tremblay & Gardner, 1996) in areas such as medicine (Horton & Switzer, 2005; Strasak, Zaman, Marinell, Pfeiffer, & Ulmer, 2007), psychology (Fabrigar, Wegener, MacCallum, & Strahan, 1999) and education (Walberg, Vukosavich, & Tsai, 1981; Zientek, Capraro, & Capraro, 2008). Such self-reflective examinations of the research community's practices are necessary to provide guidance and an informed dialogue with which to direct (or redirect) our communal path.

Another source of guidance for the research community are the positions taken by major organizations such as the American Psychological Association (APA; 2010) and the American Educational Research Association (AERA; 2006) regarding expected standards for publication. Additionally, special committees have been convened at the request of the APA to make recommendations on statistical practices in the literature (Wilkinson & the Task Force on Statistical Inference, 1999). Moreover, journal editors

This dissertation follows the style of *Educational and Psychological Measurement*.

have written recommendations for writing in their respective journals (Fan & Thompson, 2001; Thompson, 1994). These guidelines for best practices must be continuously updated because as Kieffer and colleagues explained, “the social sciences are continually evolving regarding normatively expected quantitative research practices” (Kieffer, Reese, & Thompson, 2001, p. 291).

Thus, not only is the research community interested in what gets published in terms of tabulating frequencies in particular areas of interest but much effort has been invested both individually and corporately to establish norms and expectations and to evaluate the extent to which these expectations are met (Capraro & Capraro, 2003; Fidler, 2002; Finch, Thomason, & Cumming, 2002; Vacha-Haase, Nilsson, Reetz, Lance, & Thompson, 2000). While researchers are free to choose the analytical tool with which to investigate their data, historically the majority of researchers have chosen inferential statistics. In fact in Edington’s (1964, 1974) review of the statistical practices in APA journals, he noted that 91% of the articles involved statistical inference. Statistical inference encompasses both estimation of population parameters and hypothesis testing (Siegel & Castellan, 1988). In 2007, Fidler explained that “for more than 50 years, statistical significance testing has been psychologists’ main statistical method” (Erwin, p. 1). The popularity of statistical significance testing is not restricted to psychology, as researchers have commented “null hypothesis significance testing is the workhorse of research in many disciplines, including medicine, education, ecology, economics, sociology and psychology” (Erceg-Hurn & Mirosevich, 2008, p. 591).

Despite its omnipresence, hypothesis testing, typically in the form of null hypothesis statistical significance testing (NHSST) has been a source of contention for many years.

NHSST was introduced in the early 1920s (Fisher, 1925; Fisher & Mackenzie, 1923; Hubbard, Bayarri, Berk, & Carlton, 2003; Neyman & Pearson, 1928). Not long afterward, between 1940 and 1950, a NHSST metastasis evolved (Hubbard & Ryan, 2000). Yet even as NHSST was being rapidly incorporated into researchers' analytical toolboxes, NHSST was fraught with controversy. Kaufman (1998) stated "the controversy about the use or misuse of statistical significance testinghas become the major methodological issue of our generation" (p. 1). Books have been published to explain the nature of significance testing (Mohr, 1990) by those who favor (Chow, 1996) and oppose (Harlow, Mulaik, & Steiger, 1997; Kline, 2009; Morrison & Henkel, 1973) its use. Special issues of the *Journal of Experimental Education* (Journal of Experimental Education, 1993) and *Research in the Schools* (Kaufman, 1998) have been dedicated to addressing the NHSST controversy. Even major organizations such as the AERA (2006) and the APA (2010) have taken a position on the NHSST debate. The most recent declaration given in the newest APA (2010) manual recounts,

historically, researchers in psychology have relied heavily on null hypothesis statistical significance testing (NHST) as a starting point for many (but not all) of its analytical approaches. APA stresses that NHST is but a *starting point* [emphasis added] and that additional reporting elements such as *effect sizes*, [emphasis added] confidence intervals and extensive description are needed to convey the most complete meaning of results. (p. 33)

Further it appears as if journal editors are also joining the NHSST reform movement as effect size reporting is now required for at least 24 journals (Thompson, 2008).

So it appears as though finally, the call to reform research practices is being heard. Cohen knowingly advised researchers that changing research practice takes time. He commented, “ if you publish something that you think is really good, and a year or a decade or two go by and hardly anyone seems to have taken notice... take heart” (Cohen, 1990, p. 1311). Thus it is necessary to systematically evaluate the present state of our research practices in light of our goals to be able to continually recalibrate our intended direction. Indeed, it is only when we hold the mirror up to ourselves and purposely reflect on where we are at the present time that we can begin to consider where we would like to be. To nonsensically consider where we would like to be without considering where we are now, would be like asking for directions without knowing our present location!

Effect Sizes

In 2002, the National Research Council’s *Scientific Research in Education* emphatically upheld the position of education as a scientific endeavor (Shavelson & Towne). As such, it makes sense that educational researchers, like scientific researchers, are interested in the accumulation, synthesis and integration of new knowledge in light of previous findings. Educational research progresses as we build upon the contextualization of individual findings in light of other relevant literature. In other words, “scientific knowledge advances when findings are reproduced in a range of times

and places and when findings are integrated and synthesized” (Shavelson & Towne, 2002, p. 4).

One way to integrate findings across the literature is the comparison of effect sizes across studies. Thompson identified this philosophy of meta-analytic thinking as “both (a) the prospective formulation of study expectations and design by explicitly invoking prior effect sizes and (b) the retrospective interpretation of new results, once they are in hand, via explicit, direct comparison with the prior effect sizes in the related literature”(2002, p. 28). With the vast amount of studies available, “meta-analytic thinking” is a philosophy that should be adopted by all researchers in order to make better sense of studies’ findings. Several professional entities have spoken on the issue of effect sizes. The Task Force on Statistical Inference (TFSI) urged researchers to “*always* [emphasis added] present effect sizes for primary outcomes” (Wilkinson & the Task Force on Statistical Inference, 1999). The language of the APA was not as strong, but still specified that “it is almost always necessary to include some index of effect size or strength of relationship in your Results section” (2010, pg. 34). However, the APA also stated that “complete reporting of all tested hypotheses and estimates of appropriate effect sizes and confidence intervals are the *minimum expectations* [emphasis added] for all APA journals” (American Psychological Association, 2010, p. 33).

One of the problems with effect size reporting may be that authors do not recognize effect sizes produced in their own analyses. An analysis of statistical techniques in the *American Educational Research Journal (AERJ)* and *Journal of Counseling Psychology (JCP)* reported “some authors routinely report bivariate and

multiple correlation coefficients, without recognizing these as effect-size indices and without interpreting them in relation to related previous effect-size reports” (Kieffer et al., 2001, p. 304). These findings are disconcerting especially because the APA *Publication Manual* (2001) specifically provides examples of common effect size estimates, such as Pearson r^2 and multiple correlation coefficient (R^2) and numerous scholars have written about effect sizes (Grissom & Kim, 2005; Kirk, 1996; Snyder & Lawson, 1993; Thompson, 2002).

Another concern regarding effect sizes is the lack of recognition of the relationship among techniques in the General Linear Model (GLM). Understanding that all analyses subsumed under the GLM should bring to mind three overarching concepts: all analyses (1) are correlational, (2) apply weights to measured variables to yield latent variables and (3) yield variance accounted for effect sizes analogous to r^2 (Thompson, 2006). As previously noted correlational techniques, including the Pearson r , are readily reported in the literature. What is not often recognized is that the Pearson r^2 , like R^2 , is positively biased. While previous researchers have addressed the connections between the GLM, understanding that like the R^2 , the r^2 is also positively biased (Wang & Thompson, 2007), an empirical review of all available correction formulas has not been previously reported. The present document will address this gap in the literature.

Another effect size measure, η^2 , is a variance accounted for effect size in ANOVA analogous to the r^2 in correlation (Grissom & Kim, 2005). ANOVA is also a commonly used statistical technique (Zientek et al., 2008). Almost since the time ANOVA was first introduced (Fisher & Mackenzie, 1923) much attention has been

given to understanding the results of ANOVA under assumption violations (Pearson, 1931). Because of the previously mentioned statistical significance obsession, however, violations of assumptions in ANOVA have focused primarily on the F test. Therefore, the behaviors of effect sizes under ANOVA assumption violations have largely been undetermined. The present document will empirically investigate the robustness of η^2 , ω^2 , and ϵ^2 under assumption violations.

Organization of Document

The present document is divided in five distinct sections. It should be noted that except for the first and last section, the sections are written as individual manuscripts planned for publication in peer-reviewed journals. Below is a description of each of the sections as conceptualized:

- The first section is an introductory section that presents a brief overview of the topics to be examined as well as a theoretical rationale for the individual studies.
- The second section presents a detailed literature review of statistical technique usage in the *AERJ* and select *APA* and *ACA* journals. In addition, the manner in which researchers have conceptualized and operationalized statistical technique reviews is discussed. This second section represents the first journal article.
- The third section presents findings from a Monte Carlo simulation to determine the best correction formula for Pearson r^2 . All available correction formulas to minimize the positive bias present in r^2 are empirically

investigated. Syntax is given to facilitate incorporation of the best correction formula into research practice. This third section represents the second journal article.

- The fourth section reports the results of a Monte Carlo simulation that empirically compares the robustness of a statistical significance test (overall F test) with estimates of practical significance (η^2 , ω^2 , and ϵ^2). This fourth section represents the third journal article.
- The fifth and final section is the concluding section that explicitly connects the findings from the three manuscript sections into a coherent, succinct conclusion for the entire project.

REVIEW OF REVIEWS OF STATISTICAL TECHNIQUES USED IN EDUCATION AND PSYCHOLOGY

Regardless of the discipline being taught, a course designer must decide what topics will and will not be covered in a particular course. The topics chosen reflect a prioritization of the myriad topics that could be included. Similarly, when an author decides to write a book, especially a fundamentals or introductory text, there is an understanding and expectation that certain key topics will be addressed within that text. Selected topics should provide maximum future utility and reflect current best practices. Arguably, this is even more important in applied research areas such as educational research. Our schools are the vehicle with which our society trains its populace in the “important skills, including how to read, think, communicate, inquire, study and ultimately to behave in a way that embodies the best values of the society” (Hlebowitsh, 2005, p. 43).

The topics covered in methodological courses provide researchers with the fundamental competencies necessary for scholarly investigations. As Capraro and Thompson (2008) pointed out methodological training in educational research is extremely important and “affects all educational research, its quality, and its impact on the field” (p. 247). Yet Thompson (1998) noted that doctoral curricula “seemingly have less and less room for quantitative statistics and measurement content, even while our knowledge base in these areas is burgeoning” (p. 2). An investigation of educational doctoral curricular requirements revealed that on average only 2.6 (SD=2.2) quantitative

courses were required (Capraro & Thompson, 2008). Acknowledging both the importance of educational research and the limited amount of methodological training that doctoral students receive, authors and course designers should base their selection of statistical topics on sound empirical evidence regarding what educational researchers need to know in order to be able to successfully engage in and contribute to the field.

Arguably, the methodological training requirements of doctoral curricula should be of interest to *all* consumers and producers of research in any discipline. As noted by Aiken and her colleagues, “deficiencies in quantitative and methodological training do have negative implications for the progress of substantive areas” (Aiken, West, Sechrest, & Reno, 1990, pg. 731). In psychology programs, the median statistics and measurement course requirement for doctoral students was 1.2 years (Aiken, West, & Millsap, 2008). In a similar study modeled after Aiken et al. (1990), researchers noted that while 96% of undergraduate psychology majors were required to take a statistics course, the emphasis was on “traditional approaches to analysis with relatively minimal change in response to themes and advances in the field” (Friedrich & Buday, 2000, p. 255).

Over 40 years ago, Edington (1964) recognized the need to understand what students need to know in order to “interpret the literature and to show what statistics psychologists have found useful” (p. 202). Since that time, numerous authors (Bangert & Baumberger, 2005; Edington, 1974; Elmore & Woehlke, 1988, 1996, 1998; Goodwin & Goodwin, 1985a; 1985b; Keiffer, Reese, & Thompson, 2001; Willson, 1980) have periodically assessed published literature to understand the present state of what students

need to know and be able to do in statistics. Yet, if a field is static, there is no need to continually update information regarding the selection of important topics. Indeed, once a clear understanding of what topics are important is agreed upon, it would be ridiculous to spend valuable time and effort continuously reassessing what is already known! But the field of statistics is not static, as articulated by Keselman and colleagues (1998),

improvements in statistical procedures occur on a regular basis. In particular, applied statisticians have devoted a great deal of effort to understanding the operating characteristics of statistical procedures when the distributional assumptions that underlie a particular procedure are not likely to be satisfied.

It is common knowledge that, under certain data-analytic conditions, statistical procedures will not produce valid results. The applied researcher who routinely adopts a traditional procedure without giving thought to its associated assumptions may unwittingly be filling the literature with nonreplicable results.

(p. 351)

Thus it is necessary to systematically evaluate the present state of our research practices in light of recent methodological developments not only for the sake of training future scholars, but also as a foundation upon which recommendations regarding best data analytic practices can be made.

Purpose

The purpose of the present study is to provide an historical account and synthesis of which statistical techniques were most prevalent in the fields of education and psychology. Because the *AERJ* focuses on understanding and/or improvement of

educational processes and outcomes and has consistently been reviewed regarding statistical technique usage across the years; *AERJ* was chosen as my educational research focus. To examine a parallel historical process in psychology, select *APA* and *ACA* journals were also examined. My primary research question was, how have the proportion of statistical techniques in *AERJ* varied across the years from 1969-1997? My second research question, similarly asks, how have the proportion of statistical techniques in select *APA* and *ACA* journals varied across the years from 1948-2001? Finally, I considered whether similar trends were evident across both the educational and psychological literature.

Methods

Criteria for Study Inclusion

I systematically searched *ERIC (EBSCO)*, *CSA*, *Google Scholar*, *PsycInfo*, and *Wilson Web* for the keywords listed below,

statistical technique, statistical method, statistical analysis, statistical procedure, research technique, research method, literature review, critical review, quantitative method, quantitative procedure, quantitative technique, quantitative research, review of techniques, review of methods, review of procedures, published, journal, research, publications, literature, *AERJ*, education, educational, psychology, psychological, social science.

For a study to be included, the article had to review statistical techniques in either *AERJ* or a psychology journal across more than two years. At minimum frequency counts or percentages had to be provided per technique. Preference was given to articles that

provided counts or percentages by coded year. Articles were excluded if they selectively coded only one technique per article (cf. Emmons, Stallings, & Layne, 1990) as this excluded other techniques that may have been present in the article. This was a necessary exclusion as other review articles coded multiple techniques per article therefore selecting a single technique per article coding scheme would not have been comparable to reviews that coded all given techniques per article. Next, reference sections of identified studies from the electronic database were reviewed to search for additional studies. This search produced six articles reviewing *AERJ* from 1969 to 1997 and five articles reviewing the psychological literature from 1948 to 2001, resulting in a total number of 17,698 techniques recorded from the 12,012 articles reviewed.

After eligible studies were reviewed, it became apparent that each had chosen different coding categories. To clarify commonalities among coding categories, an analysis of the coding schemes across studies was conducted. After common techniques were identified, techniques were recorded for all eligible studies by coded year (i.e., the publication year of the articles that were coded). Frequencies and percentages by technique were calculated for all studies. Calculations were necessary when either a particular study disaggregated a technique into subcategories, or if proportions instead of frequencies were provided. Once frequency counts were gathered for all reviewed studies the percentage of techniques per article were tabulated.

Identification of Eligible Techniques

Because the variables of interest to the present study are frequency counts and percentages of common univariate and multivariate techniques that had been catalogued by authors across the years in educational and psychological journals it was necessary to understand which categories of techniques were coded across the years. Table 1 displays all the categories evaluated by the authors across all the articles selected for inclusion. The statistical techniques that were most consistently reviewed across all studies were non-parametric techniques (including χ^2), t-test, regression, correlation, ANOVA, ANCOVA, factor analysis and cluster analysis. Because ANOVA and ANCOVA were treated separately in some studies and together for others, the decision was made to combine the two categories. Similarly because factor analysis and cluster analysis were coded separately in some studies and together for others, the decision was made to combine the two categories. Thus the six categories that were analyzed across the years were Nonparametric (including χ^2), *t* test, correlation, regression, ANOVA/ANCOVA, and factor/cluster analysis. Among the statistical techniques chosen for further analyses, regression is the oldest technique having been introduced in 1885 by Sir Francis Galton (David, 1995) whose notable academic offspring was Karl Pearson.

Table 1
Patterns of Statistical Technique Coding Across Articles

Statistical Technique	1,3	11	10	12	4	5	6,7,8	9
ANCOVA/ one-way ANCOVA/ factorial ANCOVA		x	x	x	x	x		x
ANOVA/ one way ANOVA/ factorial ANOVA		x	x	x	x	x		x
ANOVA/ANCOVA/ repeated measures ANOVA/ANCOVA	x		x				x	
Bayesian							x	
canonical correlation/ canonical R analysis		x	x		x	x		x
CFA								x
chi square	x	x	x	x	x	x		x
cluster analysis			x	x	x	x		x
confidence intervals								x
correlation/Pearson correlation/ other correlational/ bivariate correlation/ part/partial correlation/ simple correlation	x	x	x	x	x	x	x	x
DDA								x
descriptive		x	x	x	x	x	x	
discriminant analysis/ discriminant		x	x	x	x	x		x
EFA								x
effect sizes								x
factor analysis	x	x	x	x	x	x		x
factor/cluster							x	
graphic methods/ graphics/ interaction plot/ scattergram/ box and whisker plot/ scree							x	x
Guttman scaling					x			
HLM/ hierarchical linear modeling			x					x
interpret beta								x
interpret hit rates								x
interpret r_s								x
interpret std coefs								x
jackknife/ internal replicability/ cross validation/ bootstrap					x			x
latent partition analysis					x			
LISREL							x	

Table 1 continued

Statistical Technique	1,3	11	10	12	4	5	6,7,8	9
Logistic/ logistic regression			x					x
MANCOVA/ one way/ factorial / repeated measures MANCOVA			x	x	x			x
MANOVA/ one way/ factorial				x	x			x
MANOVA/MANCOVA/ one-way MANOVA/MANCOVA/ factorial MANOVA/MANCOVA		x	x			x		
median number of variables								x
median sample size								x
meta-analysis/ synthesis						x	x	x
multi-dimensional scaling			x		x			x
multiple comparisons					x			
multiple regression/ multiple correlation/regression/ classic regression/ multiple linear regression		x	x	x	x	x	x	x
multivariate							x	x
new non parametric/ nonparametrics/ other nonparametric	x	x	x		x	x	x	x
other		x				x		
path analysis		x			x	x		x
PDA								x
planned orthogonal comparisons /post-hoc multiple comparisons / trend analysis/ planned comparison/ post-hoc univariate/ post-hoc contrast		x	x			x		x
power					x			
Psychometric theory							x	
Qualitative							x	
reliability/ score reliability/ own data/ previous study/ not reported						x		x
secondary analysis					x			
Simulation							x	
stepwise								x
structural analysis/ SEM/ structural equation modeling			x	x	x			x
study designs								x
survival rate					x			
t test/ independent t test/ dependent t test	x	x	x	x		x	x	x

Table 1 continued

Statistical Technique	1,3	11	10	12	4	5	6,7,8	9
test variance homogeneity/ test homogeneity of regression								x
time series					x			
univariate validity						x		x
<i>Note.</i> Study ID 1, Edington, 1964; Study ID 3, Edington, 1974; Study ID 4, Willson, 1980; Study ID 5, Goodwin and Goodwin; 1985b, Study ID 6, Elmore and Woehlke, 1988; Study ID 7, Elmore and Woehlke, 1996; Study ID 8, Elmore and Woehlke, 1998; Study ID 9, Kieffer, Reese, and Thompson, 2001; Study ID 10, Bangert and Baumberger, 2005; Study ID 11, Goodwin and Goodwin, 1985a; Study ID 12, Schinka, Lalone and Broeckel, 1997.								

Table 2 lists the statistical techniques in order of introduction into the literature as identified by David (1995). These six categories combined make up the majority of the techniques that have been coded, $M=75.1\%$, 95% CI [71.47, 78.67]. Interestingly, most of these techniques are also typically identified as part of doctoral students' introductory statistics course sequence (Aiken et al., 2008).

Results

In Education

Between 1969 and 1997, authors that reviewed *AERJ* recorded a total of 2,249 statistical techniques in the 1,414 articles reviewed. Table 3 details the number of articles using particular types of statistical techniques expressed as a percentage of the number of articles reviewed for that particular year. For example, Nonpara % was calculated as the sum total of nonparametric techniques coded by year by study divided by the number of articles coded by year by study. Note that Study 4 did not code the t test; therefore there are no values available for the t test for that particular time period.

Visual analysis of these data suggests much variability from year to year for any one of the techniques. Further inspection reveals that there was more than one researcher that reviewed *AERJ* articles for a particular year (for example beginning with

Table 2
Introduction of Statistical Term by Year and Author

Statistical Technique	Year	Author Introducing the Term
regression	1885	Francis Galton
correlation	1888	Francis Galton
chi square (χ^2)	1900	Karl Person
ANOVA	1918	Ronald Aylmer Fisher
<i>t</i> test	1924	Ronald Aylmer Fisher ^a
ANCOVA	1931	Arthur L. Bailey
factor analysis	1934	George Waddell Snedecor ^b
cluster analysis	1939	Robert C. Tryon
nonparametric	1942	Jacob Wolfowitz

^aconceptually attributed to William S. Gosset. ^bconceptually attributed to Charles Spearman (1904) and Karl Pearson (1901).

1979 through 1983 and again from 1988 through 1997). However even though the same journal was coded there is generally not agreement on the type of techniques or even in the number of articles reviewed for coding. For example, if you look at 1994 study eight

Table 3
 Percentage of Techniques by Articles per Year in AERJ, 1969 – 1997

Study ID	Yr	ArtCod ^a	Nonpara% ^b	<i>t</i> test% ^c	Corr% ^d	Reg% ^e	OVA% ^f	Fac% ^g
4	1969	27	19		41	30	59	15
4	1970	33	6		36	18	52	18
4	1971	32	9		25	13	69	22
4	1972	40	13		23	8	65	13
4	1973	20	0		30	15	45	30
4	1974	19	11		11	26	58	16
4	1975	23	4		22	13	57	30
4	1976	17	0		24	29	53	18
4	1977	26	4		27	19	54	15
4	1978	43	12		28	19	60	5
6	1978	43	14	12	23	21	35	9
5	1979	32	16	13	16	28	31	9
6	1979	30	7	13	10	13	23	7
5	1980	34	12	21	38	38	47	9
6	1980	34	9	15	9	24	29	9
5	1981	36	11	11	22	42	28	8
6	1981	36	3	6	0	39	22	11
5	1982	41	22	7	44	29	51	7
6	1982	41	7	5	5	24	27	5
5	1983	46	15	9	28	26	52	7
6	1983	46	9	4	9	17	28	9
6	1984	53	13	9	8	23	42	2
6	1985	38	16	13	26	26	47	13
6	1986	44	25	18	16	30	45	9
6	1987	31	10	13	13	23	42	3
8	1988	28	0	14	25	11	36	0
9	1988	21	0	10	24	29	52	14
8	1989	17	18	24	6	18	41	6
9	1989	29	14	17	21	21	31	10
8	1990	35	0	11	14	6	23	6
9	1990	24	8	25	29	33	50	4
8	1991	38	11	8	13	11	29	0
9	1991	37	11	14	24	24	35	14

Table 3 continued

Study ID	Yr	ArtCod ^a	Nonpara% ^b	<i>t</i> test% ^c	Corr% ^d	Reg% ^e	OVA% ^f	Fac% ^g
8	1992	37	8	0	5	8	22	5
9	1992	22	18	14	14	23	50	14
8	1993	31	6	6	6	10	23	10
9	1993	22	0	9	5	23	41	14
8	1994	34	15	9	6	6	21	15
9	1994	23	13	22	26	13	39	17
8	1995	25	8	8	20	32	24	20
9	1995	20	30	20	40	15	30	15
8	1996	29	0	7	10	21	17	3
9	1996	29	0	7	14	14	48	0
8	1997	24	8	0	0	13	25	21
9	1997	24	4	4	25	33	29	29

Note. Study ID 4, Willson, 1980; Study ID 5, Goodwin and Goodwin; 1985b, Study ID 6, Elmore and Woehlke, 1988; Study ID 8, Elmore and Woehlke, 1998; Study ID 9, Kieffer, Reese, and Thompson, 2001.

^aThe number of articles coded. ^bProportion of the number nonparametric and χ^2 techniques coded per article. ^cProportion of the number of *t* tests coded per article. ^dProportion of the number of correlation techniques coded per article. ^eProportion of the number of regression techniques coded per article.

^fProportion of the number of ANOVA and ANCOVA techniques coded per article. ^gProportion of the number of factor analysis and cluster analysis techniques coded per article.

coded 34 articles whereas study 9 for that same year coded 23 articles. Even when there was agreement on the number of articles coded as in 1980 where both study five and study six agreed that there were 34 articles in *AERJ* to be coded; only the percentage of factor/cluster is consistent across both studies.

To understand why *AERJ* articles were not being coded similarly across the years, it was necessary to be able to compare across studies the process by which researchers reviewed articles. The idiosyncratic nature of the author's review process is reported in Table 4.

Table 4 shows that across the studies that reviewed *AERJ* articles, all journals for a particular year were reviewed with the exception of study nine which included only quantitative articles in their review. Additionally, all *AERJ* article reviewers explicitly excluded book reviews. Some also excluded simulations meeting notices and theoretical comments. Additionally, some authors (Elmore & Woehlke, 1988, 1996, 1998) included qualitative techniques as a category.

Table 4
How Studies Varied in Their Review of Articles

Study ID	4	5	6	8	9
Published	1980	1985	1988	1998	2001
Years Coded	1969-1978	1979-1983	1978-1987	1988-1997	1988-1997
Articles Included	all	all	all	all	quantitative
What Was Not Included?	simulation, contrivance, expository, statistical derivation, book review	book reviews	book reviews, annual meeting notices, and directories	book reviews, annual meeting notices, and directories	expository, theoretical, comment, and book reviews
Multiple Techniques Coded Per Article	yes	yes	yes	yes	yes
Description of Categories	no	yes, but tabulation given only for major techniques	no	yes	no, but categories in table detail techniques
Number of Categories of Coded Techniques	24	27	14	16	67, includes stat methods and techniques

Table 4 continued

Study ID	4	5	6	8	9
Total Articles	280	189	396	298	251
Total Techniques	481	438	549	414	368
Number of Coders	1	2	2	2	2
Interrater Reliability	no	yes, 10% random sample stratified by year, 93% agreement	first years coded and recoded	authors consulted each other on any questionable procedures	yes, 5% random sample stratified by year, greater than 90% agreement
Intrarater Reliability	yes, 10% random sample stratified by year, 91% agreement	no	no	no	no
Journal Published in	<i>Educational Researcher</i>	<i>Educational Researcher</i>	<i>Educational Researcher</i>	AERA Annual Meeting	<i>Journal of Experimental Education</i>
Reported	frequencies and percentages	frequencies and percentages	frequencies	frequencies	frequencies and percentages
Additional Analysis	χ^2 on agricultural versus biological techniques	χ^2 on basic, intermediate and advanced	no	no	box plot for effect sizes

Note. Study ID 4, Willson, 1980; Study ID 5, Goodwin and Goodwin; 1985b, Study ID 6, Elmore and Woehlke, 1988; Study ID 8, Elmore and Woehlke, 1998; Study ID 9, Kieffer, Reese, and Thompson, 2001.

For all the *AERJ* review studies except for one, there were two coders. Some of the authors were explicit in their description of how they established reliability in their coding while other authors were less forthcoming. For those studies that did indicate

their method of reliability, interrater and intrarater reliabilities were established using a 5% to 10% random sample stratified by year.

Three out of the six reviews which coded *AERJ* were published in *Educational Researcher*. Each of the studies reported frequencies of statistical techniques and three out of the five also reported percentages. The number of articles coded ranged from 189 to 396. The number of statistical techniques coded ranged from 368 to 549.

Perhaps what was most variable across studies was the number of categories of statistical techniques. The least number of categories of statistical techniques coded by study six was 14; the greatest number of categories of statistical techniques coded was 67 by study nine.

Current AERA guidelines for reporting encourage transparency of the research process and state that authors “should make it possible to follow the course of decisions about the pattern descriptions, claims, and interpretations from the beginning to the end of the analysis process” (American Educational Research Association, 2006, p. 38).

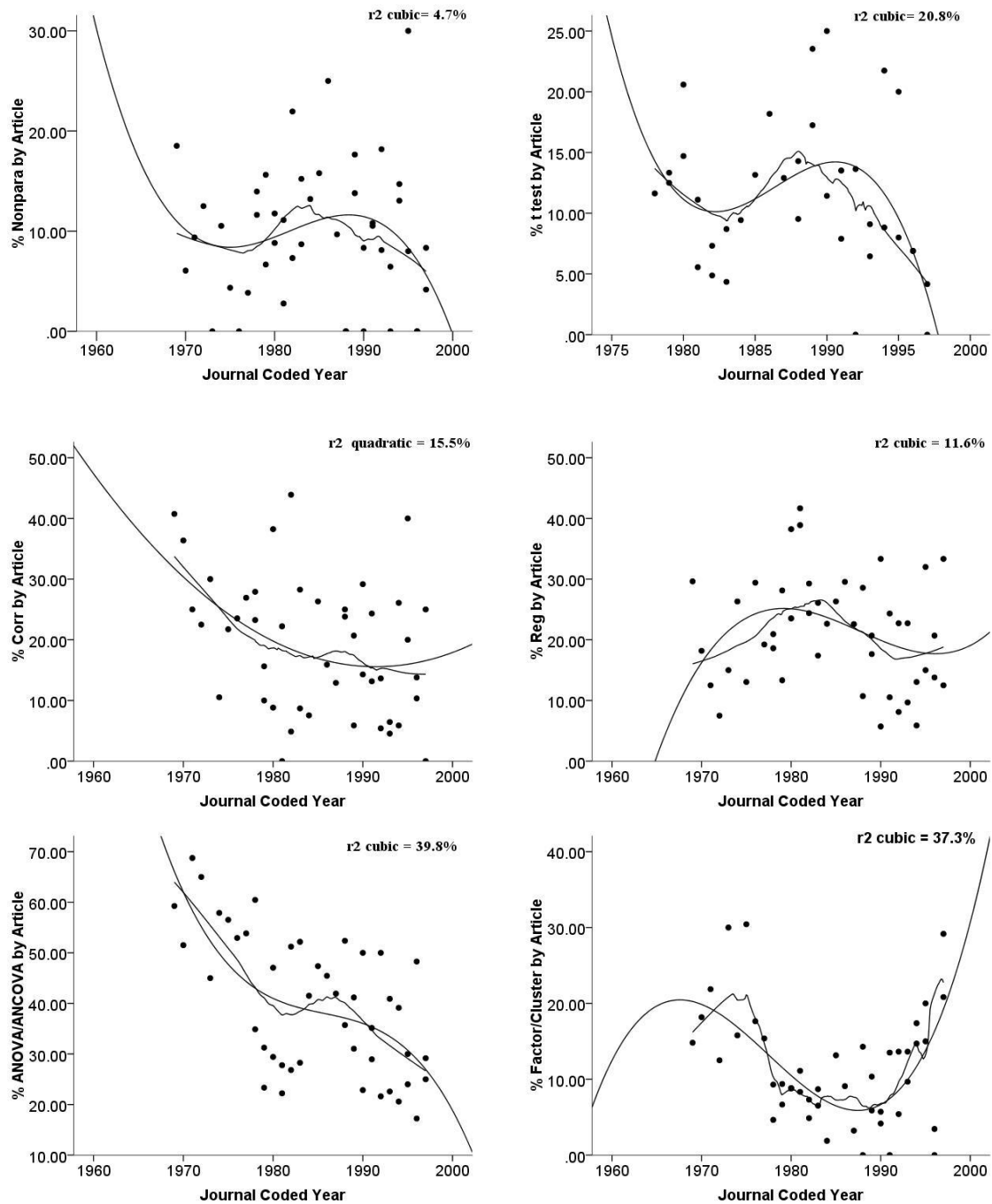
While the majority of these studies were published in *Educational Researcher*, an AERA journal; few researchers explicitly detailed what statistical techniques were included in which category. Because the number of categories varied greatly, and there was not a clear description across all studies detailing what was included and what was not included in each category, it is easy to see why there are differences within the same journal within the same year across categories. Nonetheless line graphs drawn with coded year as the independent variable and percentage of statistical technique as the criterion variable across all *AERJ* studies revealed that while the absolute magnitude of

the percentages may have differed across the same coded year for the same journal due to coding differences, the general pattern produced when studies overlapped was consistent.

While these tables provide yearly proportions of the types of statistical techniques present per article, the general trends may not be immediately apparent in tabular form. The scatterplots in Figure 1 provide a pictorial representation of the general trends across time for each of the statistical techniques. Both a loess line and an either cubic or quadratic regression line has been imposed on these scatterplots. Loess lines are localized polynomial regressions or locally weighted regressions that give “insight into the behavior of the data and help us choose parametric models” (Cleveland & Devlin, 1988, p. 596). The inclusion of the loess line and the scatter plot conveys more information to the reader than a simple trend line (Wilkinson & the Taskforce on Statistical Inference, 1999).

Generally, you can see that nonparametric techniques in *AERJ* are quite variable only 4.7% of this variability is accounted for by the coded year. On the other hand, *t* test while still showing quite a bit of variability, appear to be taking more of a downward trend in the mid to late 1990s ($r^2 = 20.8\%$). Generally the trend for correlations, while quite variable, appears to have decreased as well since the 1970s ($r^2 = 15.5\%$). Regression on the other hand may be on the rise ($r^2 = 11.6\%$). ANOVA/ANCOVA techniques have the highest r^2 value of the six studied statistical techniques (39.8%), showing a marked decreased from the 1970s to the 1990s. Finally, factor/cluster

Figure 1. Statistical Technique Reporting Trends in AERJ from 1969 to 1997 as a Percentage of Statistical Technique per Article



techniques seemed to reach a low around the mid to late 1980s, but appear to be increasing in frequency in the 1990s ($r^2 = 37.3\%$).

In Psychology

Table 5 provides the percentages of statistical techniques coded per article in psychology journals by year from 1948 to 2001. Note that studies one and three did not code for regression; therefore there are no values available from 1948 to 1972. For this table there is an overlap in the years coded by study for study 9 and 10 and for study 10 and 12. However, study 10 and 12 coded across multiple years and did not disaggregate their information by year. Therefore the median year was recorded for the aggregated results that were coded. In other words, the percentages represented in study 10 and 12 represent an aggregate of 11 and 5 years respectively recorded as coding for the median year, 1995 and 1992, respectively.

Visual inspection of the data presented reveal that the number of statistical techniques coded by Edington is much larger than any of the number of statistical techniques coded across any one year from the other studies. This is because Edington did not code a single journal but rather coded seven journals together. So each year noted in Edington's 'Yr' column actually is an aggregate of seven journals.

Scatterplots are presented for the psychology reviews as well. The patterns are generally more pronounced than they were in education. One reason may be that there simply exists more variability in *AERJ* from year to year. Another plausible reason is that for some of the years several journals are represented within one coded year

Table 5
 Percentage of Statistical Techniques by Articles per Year in Psychology Journals, 1948 – 2001

StudyID	Yr	ArtCod	Nonpara%	<i>t</i> test%	Corr%	Reg%	OVA%	Fac%
1	1948	204	13	51	42		11	1
1	1950	310	22	47	38		16	3
1	1952	458	23	44	32		24	4
1	1954	529	27	34	33		34	3
1	1956	587	32	29	32		38	3
1	1958	689	33	26	29		41	4
1	1960	593	32	23	28		42	5
1	1962	775	30	19	24		55	5
3	1964	674	36	18	20		58	4
3	1966	925	35	18	20		62	2
3	1968	1128	33	13	21		69	2
3	1970	1170	26	13	25		70	3
3	1972	1195	27	12	25		71	4
11	1979	31	16	32	42	10	65	16
11	1980	32	16	25	63	22	78	6
11	1981	30	10	33	60	17	87	17
11	1982	29	14	31	69	45	62	17
11	1983	28	11	14	57	25	89	7
9	1988	36	19	11	42	8	33	14
9	1989	64	9	13	48	19	31	13
9	1990	63	8	6	33	22	35	13
9	1991	51	6	2	37	27	2	2
9	1992	59	10	5	29	19	34	12
12	1992	449	22	24	50	12	34	21
9	1993	46	24	26	41	35	52	4
9	1994	42	29	0	55	26	21	14
9	1995	59	24	0	54	31	17	17
10	1995	256	15	13	40	26	45	6
9	1996	42	24	0	71	29	45	12
9	1997	44	23	0	55	32	41	27

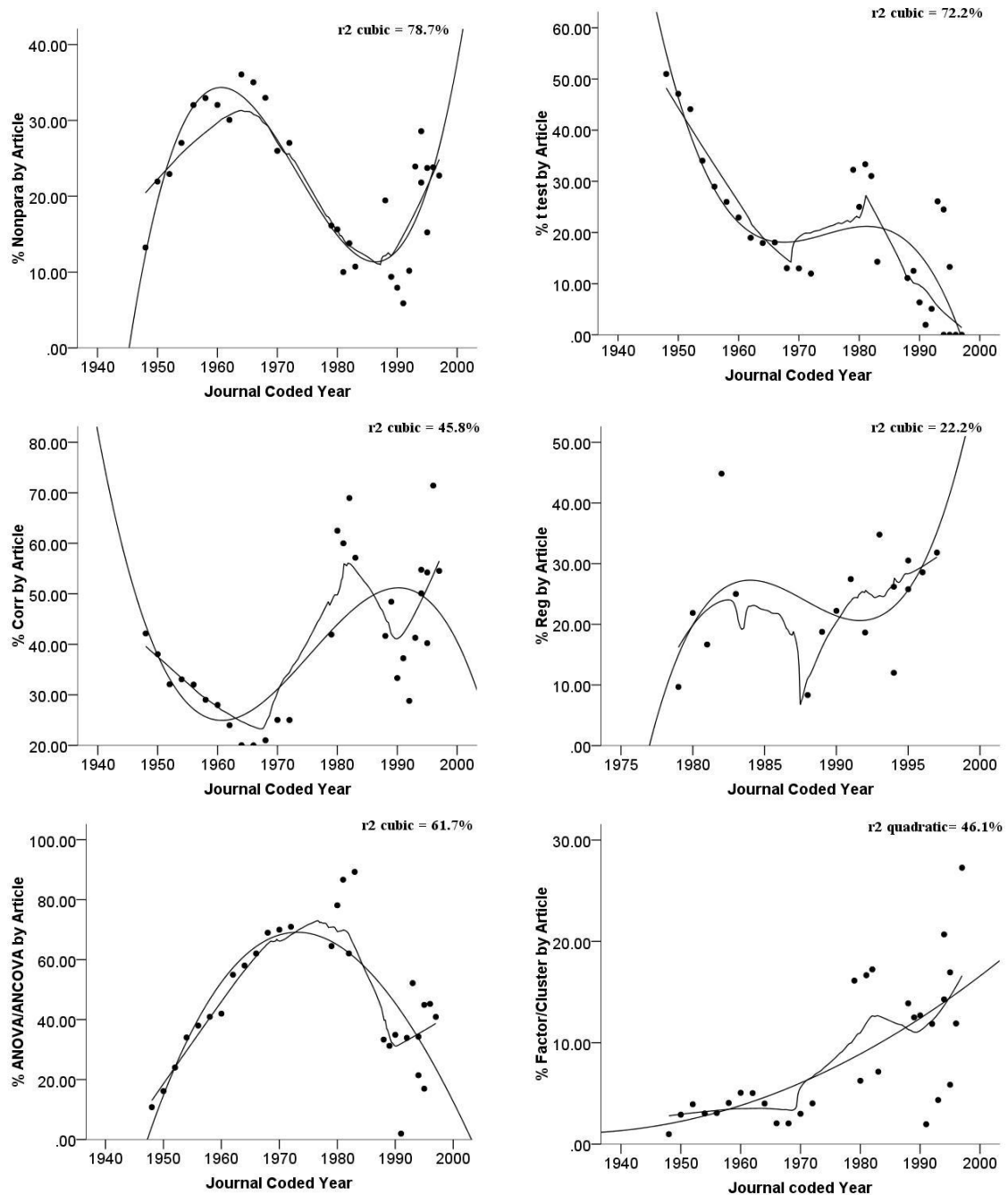
Note. Study ID 1, Edington, 1964; Study ID 3, Edington, 1974; Study ID 9, Kieffer, Reese, and Thompson, 2001; Study ID 10, Bangert and Baumberger, 2005; Study ID 11, Goodwin and Goodwin, 1985a; Study ID 12, Schinka, LaLone and Broeckel, 1997.

(viz., Edington, 1964; Edington, 1974) as mentioned above. In any case, the r^2 values are much higher for the psychology journals than the education journal indicating that a larger amount of the variability that is present across the years in the proportion of statistical technique usage per article can be accounted for by the coding year in psychology than in education.

Figure 2 graphically displays that nonparametric techniques reached a peak in the 1960s then began to decrease until more recently when they appear to be increasing in frequency again ($r^2 = 78.7\%$). Interestingly, the bootstrap was introduced by Efron (1979) in the late 1970s, and may be part of the reason for the more recent upswing in the use of nonparametric techniques along with the use of χ^2 tests as indices of overall model fit in structural equation modeling.

On the other hand, t -test usage in the psychological literature appears to have been steadily decreasing since the 1940s ($r^2 = 72.2\%$). In contrast, correlational techniques that were at their lowest in the mid to late 1960s, have been seen more frequently in the literature ($r^2 = 45.8\%$). The use of regression techniques appears to be increasing ($r^2 = 22.2\%$). ANOVA techniques have a definite curvilinear relationship and appear to have been steadily decreasing in the 1990s ($r^2 = 61.7\%$). Finally, factor analysis, which Edington (1964) recommended “be taught outside the general course in

Figure 2. Statistical Technique Reporting Trends in Psychology Journals from 1948 to 2001 as a Percentage of Statistical Technique per Article



statistics” (p. 202) appears to be appearing more frequently in psychological journals ($r^2 = 46.1\%$).

Education and Psychology Results as a Whole

Interesting patterns emerged when examining reviews of education and psychology separately. Now, the focus turns to examining how the paths of education and psychology have evolved compared to each other and in aggregate. Table 6 lists the means and confidence intervals for the reviewed studies for the most recent decade

Table 6
95% Confidence Intervals for Mean Percentage of Statistical Usage per Article from 1990- 1997

Statistical Technique	Education ($n=16$)				Psychology ($n=10$)				All Journals ($n=26$)			
	95% CI				95% CI				95% CI			
	<i>M</i>	<i>SD</i>	LL	UL	<i>M</i>	<i>SD</i>	LL	UL	<i>M</i>	<i>SD</i>	LL	UL
Nonparametric	8.79	7.90	4.58	13.00	18.38	7.93	12.71	24.05	12.48	9.09	8.81	16.15
<i>t</i> test	10.22	7.15	6.41	14.03	7.73	10.16	0.45	15.00	9.26	8.33	5.90	12.62
Correlation	15.76	10.84	9.98	21.53	46.60	12.76	37.47	55.73	27.62	19.06	19.92	35.32
Regression	17.71	9.52	12.64	22.78	25.80	6.70	21.00	30.59	20.82	9.31	17.06	24.58
ANOVA/ ANCOVA	31.59	11.00	25.73	37.45	32.67	15.21	21.79	43.55	32.01	12.49	26.96	37.05
Factor Analysis/ Cluster Analysis	11.64	8.16	7.30	15.99	12.78	7.67	7.29	18.27	12.08	7.84	8.91	15.25

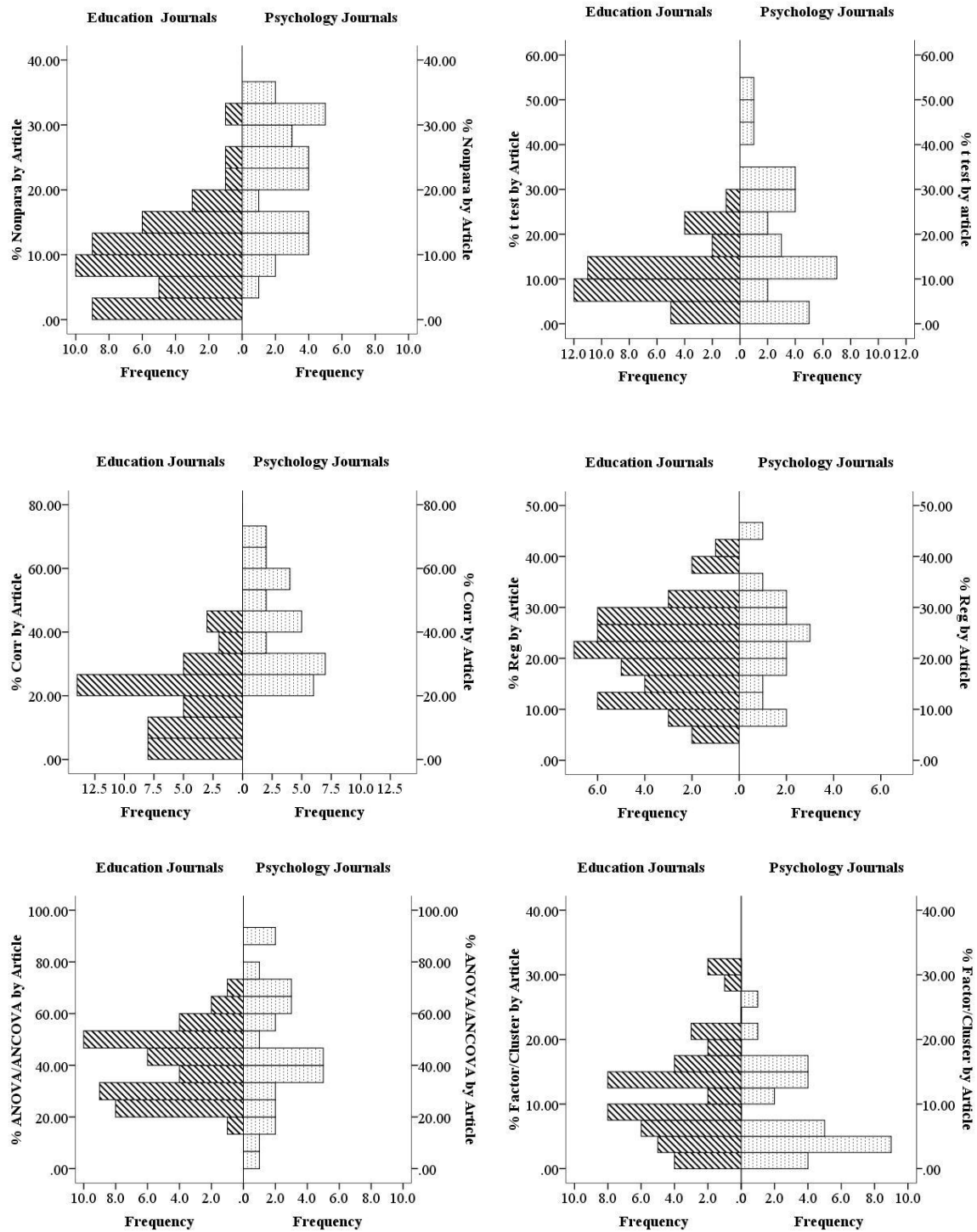
available, 1990s. Across this time period, ANOVA techniques have been most popular in education and overall. In psychology, correlation has been the most prevalent statistical technique. The second most popular statistical technique has been regression, ANOVA, and correlation in education, psychology and combined, respectively.

Closer inspection of Table 6 reveals that the standard deviations are large which reflects the variability from year to year in the percentage of statistical techniques coded per article. This variability is something that can be explored further. Figure 3 captures the general distribution in percentage of statistical techniques per article in educational and psychological journals. Whereas the frequency distribution is somewhat parallel for regression, ANOVA, and factor analysis, the distributions are more different for nonparametric, *t* test and especially correlation.

Utility of Research Findings

The present study has aggregated previous reviews of statistical techniques to represent trends of statistical technique usage across time. This information is useful to course designers as a general guide of what topics students need in order to be prepared to be a literate producer and consumer of research. Similarly the present study is useful to book authors of introductory texts, as they decide what topics would be most relevant to current and future researchers. Other audiences that may find the present document useful are methodologists, journal editors, reviewers and major organizations such as APA and AERA. These individual and corporate entities make decisions regarding analytical practices that promote good scholarship. The gold standard in decision making is evidence-based practice. Evidence-based decisions are made in light of “best

Figure 3. Statistical Technique Distributions in Psychology Journals as a Percentage of Statistical Technique per Article



available research” (American Psychological Association, 2006) together with “professional wisdom” (Whitehurst, 2002, p. slide 2). The evidence provided in the present document, together with the individual and collective professional wisdom of methodologists, journal editors, reviewers and major organizations can help direct the research community’s path towards continued improvement.

According to the *ISI Web of Knowledge*, in 2008, there were over 600 education and psychology journals with a combined total of over 30,000 articles. Truly no single study in and of itself can paint as complete a picture as an aggregate systematic review across studies. However, this does not detract from the value of single studies, as a systematic review would not be possible without previous studies. Undoubtedly quantitative reviews, such as the present document, that cumulate previous studies aid the research community by integrating previous knowledge.

Recommendations for Future Reviews of Statistical Techniques

Some recommendations to facilitate “sufficiency of warrants and ... transparency” in reporting are offered for authors of statistical technique reviews (American Educational Research Association, 2006, p. 33). First, authors should explicitly describe their coding scheme as suggested by AERA (2006) “any classification scheme should be comprehensively described and illustrated with concrete examples that represent the range of phenomena classified” (p. 36). Second, it is important to provide frequency counts or percentages, along with the number of techniques coded and the number of articles coded *by year*, as this allows for future summative reviews across studies. Future reviewers can always aggregate data, but

cannot disaggregate previously aggregated data. Similarly, the coding categories should be provided in the lowest possible unit (i.e., *more* rather than less categories) so that researchers with varied research questions can extract relevant information.

As should be evident by the results of the present study, the statistical technique of choice is not fixed nor is it necessarily the same for psychological and educational researchers. The techniques that were most used at one time may not necessarily still be the same techniques that are being used most frequently today. As such, our teaching practices, reviewing processes, and reporting recommendations need to consider where current practices are now as decisions are made regarding future data analytic directions.

CHOOSING THE BEST CORRECTION FORMULA FOR PEARSON r^2

Introduction

Education researchers are in the business of accumulating knowledge to inform educational practice, programs and policies. Yet no single study can be relied upon to unequivocally provide definitive answers based on that study's finite set of observations. New educational research findings emerge almost daily from various sources, such as books, reports, government documents, websites, and journal articles. In 2008 alone, the *ISI Web of Knowledge* reported over 100 education and educational research journals in the *Journal Citation Index*. Each journal in that year had an average of 39 articles. So in one year, there were well over 4,000 educational research articles to glean information from, notwithstanding the various other sources of education research! Surely, it is easy to recognize the wisdom of researchers that urge the research community to synthesize and integrate research findings across studies (Hunter & Schmidt, 2004; Shavelson & Towne, 2002; Thompson, 2002).

In 2002, the National Research Council's *Scientific Research in Education* emphatically upheld the position of education as a scientific endeavor (Shavelson & Towne). As such, it makes sense that education researchers, like scientific researchers, are interested in the accumulation, synthesis and integration of new knowledge in light of previous findings. Educational research progresses as we build upon the contextualization of individual findings in light of other relevant literature. In other words, "scientific knowledge advances when findings are reproduced in a range of times

and places and when findings are integrated and synthesized” (Shavelson & Towne, p. 4).

One way to integrate findings across the literature is the comparison of effect sizes across studies. Thompson identifies this philosophy of meta-analytic thinking as “both (a) the prospective formulation of study expectations and design by explicitly invoking prior effect sizes and (b) the retrospective interpretation of new results, once they are in hand, via explicit, direct comparison with the prior effect sizes in the related literature”(2002, p. 28). With the vast amount of studies available, “meta-analytic thinking” is a philosophy that should be adopted by all researchers in order to make better sense of studies’ findings. Several professional entities have spoken on the issue of effect sizes. The Task Force on Statistical Inference (TFSI) urged researchers to “Always present effect sizes for primary outcomes” (Wilkinson & the Task Force on Statistical Inference, 1999). The language of the American Psychological Association (2001) was not as strong, but still specified “it is almost always necessary to include some index of effect size or strength of relationship in your Results section” (p. 25).

One of the problems with effect size reporting may be that authors do not recognize effect sizes produced in their own analyses. An analysis of statistical techniques in the *AERJ* and the *Journal of Counseling Psychology(JCP)* reported “some authors routinely report bivariate and multiple correlation coefficients, without recognizing these as effect-size indices and without interpreting them in relation to related previous effect-size reports” (Kieffer, Reese, & Thompson, 2001, p. 304). This is disconcerting especially because APA’s *Publication Manual* (2001) specifically

provides examples of common effect size estimates, such as Pearson r^2 and multiple correlation coefficient (R^2), and numerous scholars have written about effect sizes (Grissom & Kim, 2005; Kirk, 1996; Snyder & Lawson, 1993; Thompson, 2002).

Bias in Regression

Statisticians have long recognized that effect sizes, including the multiple R^2 , which are estimated with ordinary least squares (OLS) theory produce biased estimates (Wang & Thompson, 2007; Wherry, 1931). Corrections formulas are frequently applied with multiple R^2 values to adjust for this bias. Numerous correction formulas are available (Claudy, 1978; Olkin & Pratt, 1958; Stuart, Ord, & Arnold, 1994; Wherry, 1931). One multiple R^2 formula due to Ezekiel is the default in common statistical packages (Ezekiel, 1929; Wang & Thompson, 2007; Yin & Fan, 2001) and touted as the “most widely used equation for estimating the squared population multiple correlation, given a sample multiple correlation coefficient” (Claudy, p. 596). If we recognize the links between variance-accounted-for effect sizes within the framework of the General Linear Model, we realize that Pearson’s r^2 also produces a positively biased estimate (Thompson, 2006). Olkin and Pratt proposed a formula to correct for the bias in Pearson r . However the use of correction formulas for bivariate correlations is not common in the literature.

Positively biased estimates are especially a concern at smaller sample sizes. Yet there are other factors that impact sampling error variance. When we calculate sample statistics we are working with a finite number of cases believed to be part of a larger theoretical population of interest. Recognizing that all samples have some information

that exists in the population and also some information that is not present in the population but unique to our particular sample, it is prudent to take this sampling error variance into account. Three variables affect sampling error variance, (1) sample size, (2) the number of predictor variables, and (3) effect size. As sample size increases, better estimates of the population of interest are obtained. This makes sense intuitively if you consider that as you increase the sample size, you approach the size of the population.

To understand why the number of predictor variables affects the sampling error variance, consider that each measured variable included in a model increases the opportunity for the “weirdness of people to be manifested” (Thompson, 2006, p. 194). In other words, the more variables you measure, the more likely you are to find an extreme value that is inconsistent with the others for that particular variable for a given sample statistic. Predictor variables are theoretically or empirically selected based on their ability to explain the variability that exists in the sample data. Effect sizes provide information about the amount of variance in the criterion variable that is explained by the predictor variables. The coefficient of determination, r^2 , is one such effect size. In the case of the Pearson r , there is only one predictor. Thus, having one predictor, as opposed to more than one, minimizes sampling error variance.

To understand why population parameter effect sizes impact sampling error variance, we can consider the extreme case of $r^2=1.0$. In this case, all of the variability present in the criterion variable is explained by the predictor. Whether you have two participants or two million participants in your sample, any two cases will depict a

perfectly explained bivariate relationship, r^2 . Thus, as the coefficient of determination approaches one, smaller sample sizes are able to give remarkably accurate estimates of the effect size.

Purpose

Recently, researchers have investigated multiple R^2 correction formulas to determine their utility in correcting Pearson r^2 bias (Wang & Thompson, 2007). The present study determines, via a Monte Carlo simulation, the utility of seven correction formulas, two of which have not been studied previously with the present design features, Olkin-Pratt Pearson r and the multiple R^2 Olkin-Pratt extended formula. The Olkin-Pratt Pearson r formula was specifically designed to correct for Pearson r bias. Previous researchers have studied the efficacy of multiple R^2 correction formulas with multiple predictors (Cattin, 1980; Claudy, 1978; Huberty & Mourad, 1980; Raju, Bilgic, Edwards, & Fleer, 1999; Yin & Fan, 2001), multiple R^2 correction formulas with one predictor (Cattin, 1980; Wang & Thompson, 2007), and the Olkin-Pratt Pearson r correction formula (Zimmerman, Zumbo, & Williams, 2003). Because previous studies differ regarding which equations were studied, type of data (raw data or simulation), number of predictors, and various design features (such as values of ρ , sample size, and shape distributions), it is not possible to judge which formula provides the best unbiased correction estimate. Therefore, the primary purpose of the present study was to identify the correction formula(s) that consistently produced an unbiased estimate of coefficient of determination across the design features investigated.

R^2 and Pearson r^2 Correction Formulas

The multiple R^2 correction formulas are, perhaps not surprisingly, very similar to each other. As you can see in Table 7, all multiple R^2 formulas contain the term $(1-R^2)$ in the numerator. All multiple R^2 formulas contain the term $(n-p)$ in the denominator. All multiple R^2 formulas begin by subtracting from unity. In addition, there are pairs of formulas that are almost identical. For example, Claudy's formula and Olkin-Pratt's formula differ only in the numerator, where Claudy has $(n-4)$ while Olkin-Pratt has $(n-3)$. Olkin-Pratt's and Pratt's formulas differ only in the denominator where Olkin-Pratt has $n-p+1$, while Pratt has $n-p-2$. The most parsimonious of the formulas, Smith's and Ezekiel's differ in the numerator and denominator by -1 (Smith, $\frac{N}{N-p}$, Ezekiel, $\frac{N-1}{N-p-1}$).

In addition to providing the formulas in Table 7, I have included the Excel formulas to facilitate the use of these formulas for the interested reader. Figure 4 shows an Excel sheet with the formulas in the appropriate cells. If these formulas are simply copied and pasted into place, then the calculation will be easily performed in Excel. After the formulas are in place the researcher need only input their values of R^2 (cell B1), n (cell B2) and p (cell B3) in the shaded boxes in order to obtain corrected values in cells B5-B12.

Table 7
Six R^2 Correction Formulas and a Pearson r^2 Correction Formula with Excel Formula Syntax

Formula Name	Formula	Excel
R^2 Correction Formulas:		
Claudy (1978, p. 603)	$\hat{R}_C^2 = 1 - \frac{(n-4)(1-R^2)}{n-p-1} \left[1 + \frac{2(1-R^2)}{n-p+1} \right]$	=1-(((B\$2-4)*(1-B\$1))/(B\$2-B\$3-1))*(1+(2*(1-B\$1))/(B\$2-B\$3+1))
Ezekiel (1929; 1930, p. 121)	$\hat{R}_E^2 = 1 - \frac{n-1}{n-p-1} (1 - R^2)$	=1-((B\$2-1)/(B\$2-B\$3-1))*(1-B\$1)
Olkin-Pratt (1958; Yin & Fan, 2001, p. 208)	$\hat{R}_{OP}^2 = 1 - \frac{(n-3)(1-R^2)}{n-p-1} \left[1 + \frac{2(1-R^2)}{n-p+1} \right]$	=B\$1-((B\$3-2)/(B\$2-B\$3-1))*(1-B\$1)-(2*(B\$2-3))/((B\$2-B\$3-1)*(B\$2-B\$3+1))*(1-B\$1)^2
Olkin-Pratt Extended (1958; Cattin, 1980, p. 409)	$\hat{R}_{OP}^2 = 1 - \frac{(n-3)(1-R^2)}{n-p-1} \left[1 + \frac{2(1-R^2)}{n-p+1} + \frac{8(1-R^2)^2}{(n-p-1)(n-p+3)} \right]$	=1-((B\$2-3)*(1-B\$1)/(B\$2-B\$3-1))*(1+((2*(1-B\$1))/(B\$2-B\$3+1))+((8*(1-B\$1)^2)/((B\$2-B\$3-1)*(B\$2-B\$3+3))))
Pratt (personal communication to E. E. Cureton, October 20, 1964, cited in Claudy, 1978, p. 597)	$\hat{R}_P^2 = 1 - \frac{(n-3)(1-R^2)}{n-p-1} \left[1 + \frac{2(1-R^2)}{n-p-2.3} \right]$	=1-(((B\$2-3)*(1-B\$1))/(B\$2-B\$3-1))*(1+(2*(1-B\$1))/(B\$2-B\$3-2.3))
Smith (Ezekiel, 1929; Yin & Fan, 2001, p. 207)	$\hat{R}_S^2 = 1 - \frac{n}{n-p} (1 - R^2)$	=1-(B\$2/(B\$2-B\$3))*(1-B\$1)
Pearson r Correction Formula:		
Olkin-Pratt Pearson (1958, p. 203)	$r_{op}^2 = \left(r * \left[1 + \frac{1-r^2}{2(n-3)} \right] \right)^2$	=((B\$1^0.5)*(1+(1-B\$1)/(2*(B\$2-3))))^2

Note. For the Excel formulas, the value for R^2 should be in cell B1, the value for the sample size (n) should be in cell B2, the value for the number of predictors (p) should be in cell B3 as in Figure 4.

Figure 4. Excel Sheet with Correction Formulas

	A	B
1	$r^2=$	0.01
2	$n=$	10
3	$p=$	1
4		
5	Adjusted Claudy	$=1-(((C\$3-4)*(1-C\$2))/(C\$3-C\$4-1))*(1+(2*(1-C\$2))/(C\$3-C\$4+1))$
6	Adjusted Ezekiel	$=1-((C\$3-1)/(C\$3-C\$4-1))*(1-C\$2)$
7	Adjusted Olkin-Pratt	$=C\$2-((C\$4-2)/(C\$3-C\$4-1))*(1-C\$2)-(2*(C\$3-3))/((C\$3-C\$4-1)*(C\$3-C\$4+1))*(1-C\$2)^2$
8	Adjusted Olkin-Pratt Extended	$=1-((C\$3-3)*(1-C\$2)/(C\$3-C\$4-1))*(1+((2*(1-C\$2))/(C\$3-C\$4+1))+((8*(1-C\$2)^2)/((C\$3-C\$4-1)*(C\$3-C\$4+3))))$
9	Adjusted Olkin-Pratt Pearson	$=((C\$2^0.5)*(1+(1-C\$2)/(2*(C\$3-3))))^2$
10	Adjusted Pratt	$=1-(((C\$3-3)*(1-C\$2))/(C\$3-C\$4-1))*(1+(2*(1-C\$2))/(C\$3-C\$4+2.3))$
11	Adjusted Smith	$=1-(C\$3/(C\$3-C\$4))*(1-C\$2)$
12	Uncorrected Wherry	$=1-((C\$3-1)/(C\$3-C\$4))*(1-C\$2)$

Note. After the correction formulas are inputted, appropriate values can be inputted in the three cells shaded in grey (B1, B2, B3). Resulting calculations will appear in cells B5-B12.

Because some researchers may be more likely to use SPSS rather than Excel for their analyses, I have also included the SPSS syntax for the correction formulas in Table 8. For this syntax to work as written, the researcher should include the following variables in their data file: RSQ (R^2), n (sample size), and p (number of predictors). After these variables are in place, the syntax will compute adjusted R^2 values as indicated. Note that in the Pearson r case of one predictor, the Wherry correction does not produce a corrected value. Inputting a 1 for the 1 predictor case in the Wherry formula, $1-((n-1)/(n-p))*(1-RSQ)$, results in $(1-((n-1)/(n-1))*(1-RSQ))$. Thus the term $(n-1)/(n-1)$ reduces to one and for this one predictor case, the Wherry formula is $1-(1)*(1-RSQ)$ which reduces to $1-1+RSQ$ which is equal to RSQ . Of course, if there is more than one predictor, a corrected value will be obtained.

Table 8
SPSS Syntax

```
set printback=listing tvars=both tnumbers=both .
list variables=all/cases=99999/format=numbered .
descriptive variables=all/statistics=all .
```

```
Comment ***** Claudy Correction ***** .
Compute RSQ_C=1-(((n-4)*(1-RSQ))/(n-p-1))*(1+(2*(1-RSQ))/(n-p+1)) .
Execute .
```

```
Comment ***** Ezekiel Correction ***** .
Compute RSQ_E=1-((n-1)/(n-p-1))*(1-RSQ) .
Execute .
```

```
Comment ***** Olkin-Pratt Correction ***** .
Compute RSQ_OP=RSQ-((p-2)/(n-p-1))*(1-RSQ)-(2*(n-3))/((n-p-1)*(n-p+1))*(1-
RSQ)**2 .
Execute .
```

```
Comment ***** Olkin-Pratt Extended Correction ***** .
Compute RSQ_OPE=1-((n-3)*(1-RSQ)/(n-p-1))*(1+((2*(1-RSQ))/(n-p+1))+((8*(1-
RSQ)**2)/((n-p-1)*(n-p+3)))) .
Execute .
```

```
Comment ***** Olkin-Pratt Pearson Correction ***** .
Compute RSQ_OPP=((RSQ**0.5)*(1+(1-RSQ)/(2*(n-3))))**2 .
Execute .
```

```
Comment ***** Pratt Correction ***** .
Compute RSQ_P=1-(((n-3)*(1-RSQ))/(n-p-1))*(1+(2*(1-RSQ))/(n-p-2.3)) .
Execute .
```

```
Comment ***** Smith Correction ***** .
Compute RSQ_S=1-(n/(n-p))*(1-RSQ) .
Execute .
```

```
Comment ***** Wherry Correction ***** .
Compute RSQ_W=1-((n-1)/(n-p))*(1-RSQ) .
Execute .
```

Note: Variables RSQ (R^2), n (sample size), and p (number of predictors) are variables in the SPSS data file.

Correction and Shrinkage

Because the Pearson r produces positively biased estimates, you should expect that corrected or adjusted r^2 values would result in a lower value than the uncorrected r^2 . This is seen in most of the corrected values for r^2 in Table 9 although in some cases the correction actually produced an even more positively biased r^2 estimate (see Claudy's negative shrinkage values below). As you can see, the differences in corrected and uncorrected r^2 values are most extreme at the smaller sample sizes and lower values of r^2 . This should not be surprising because smaller sample sizes and smaller effect sizes produce greater sampling error variances, as described previously. Because the Pearson r , which always has a single predictor, was the focus of the present study, the number of predictors was not a variable.

Methods

Using the RANNOR random number generator in SAS (version 9.1), 5000 samples were drawn from each of the 108 (6 x 3 x 6) simulation conditions (i.e., population ρ values of 0.0, 0.1, 0.3, 0.5, 0.7, and 0.9; population shapes normal, skewness = kurtosis = 1, and skewness = -1.5 with kurtosis = 3.5; ns = 10, 20, 40, 60, 100, and 200 respectively). SAS generated population data with inputted correlation and shape parameters using Vale and Maurelli's (1983) multivariate extension of Fleishman's procedure (1978). To confirm that the program was working as

Table 9
Adjusted r^2 Values for Selected Sample Sizes and Uncorrected r^2 Values

Formula Name	n	r^2	Adjusted r^2	Shrinkage %
Claudy	10	0.01	0.110	-1004.85%
	200	0.01	0.010	-2.98%
	10	0.81	0.852	-5.20%
	200	0.81	0.812	-0.19%
Ezekiel	10	0.01	-0.114	1237.50%
	200	0.01	0.005	50.00%
	10	0.81	0.786	2.93%
	200	0.81	0.809	0.12%
Olkin-Pratt	10	0.01	-0.038	477.68%
	200	0.01	0.005	47.52%
	10	0.81	0.827	-2.15%
	200	0.81	0.811	-0.07%
Olkin-Pratt Extended	10	0.01	-0.109	1185.18%
	200	0.01	0.005	49.45%
	10	0.81	0.827	-2.09%
	200	0.81	0.811	-0.07%
Olkin-Pratt Pearson	10	0.01	0.011	-14.64%
	200	0.01	0.010	-0.50%
	10	0.81	0.832	-2.73%
	200	0.81	0.811	-0.10%
Pratt	10	0.01	-0.122	1322.46%
	200	0.01	0.005	49.15%
	10	0.81	0.824	-1.77%
	200	0.81	0.811	-0.07%
Smith	10	0.01	-0.100	1100.00%
	200	0.01	0.005	49.75%
	10	0.81	0.789	2.61%
	200	0.81	0.809	0.12%

Note. Percent shrinkage is calculated as $(r^2 - \text{adjusted } r^2) / r^2$.

programmed, populations of 100,000 scores to compare the obtained parameters with the specified parameters were created. Detailed SAS programming explanations, including sample programs can be found in *SAS for Monte Carlo Studies* (Fan, Felsovalyi, Sivo, & Keenan, 2001).

In the present study six multiple R^2 formulas (Claudy, Ezekiel, Olkin-Pratt, Olkin-Pratt Extended, Pratt, and Smith) and one Pearson r formula (Olkin-Pratt Pearson r) were evaluated. In each of the 540,000 (5,000 x 108) samples (a) bias and (b) precision under each of the 108 simulation conditions were estimated. Bias was computed by subtracting the known population ρ^2 value from the sample r^2 estimate. Thus, positive bias values reflect adjusted sample r^2 values that overestimated population parameters. Of course, unbiased estimates by definition should have mean bias values of zero. Precision was defined as the standard deviation of the corresponding sample statistics (Wang & Thompson, 2007).

Results

Table 10 provides the mean bias for each of the six multiple R^2 correction formulas, the Pearson r correction formula and the Wherry correction (which for the one predictor case makes no correction) across each of the 108 simulation conditions. From the uncorrected Wherry estimates you can see that bias was consistently greatest at smaller sample sizes and smaller values of ρ^2 , as expected. For example, for a normally distributed population where $\rho^2 = 0.00$ ($\rho = 0.00$), the mean sample r^2 estimate was $\rho^2 = 0.112$ ($\rho = .335$) across the sample of 5,000. On the other hand, for the normally

Table 10
Mean Bias (Estimate - Parameter ρ^2) Across 5,000 Samples Drawn in Each of 108
Simulation Conditions

ρ^2	Shape	n	Corrected r^2							Uncorrected r^2
			Claudy	Ezekiel	Olkin-Pratt	Olkin-Pratt Extended	Olkin-Pratt Pearson	Pratt	Smith	Wherry
0.00	1	10	0.214	0.002	0.082	0.028	0.124	0.013	0.014	0.112
0.00	2	10	0.213	0.001	0.082	0.028	0.124	0.013	0.014	0.112
0.00	3	10	0.211	-0.002	0.079	0.025	0.121	0.009	0.011	0.110
0.01	1	10	0.210	-0.001	0.080	0.027	0.122	0.012	0.011	0.109
0.01	2	10	0.208	-0.003	0.078	0.024	0.119	0.009	0.009	0.107
0.01	3	10	0.277	0.078	0.158	0.113	0.196	0.098	0.089	0.179
0.09	1	10	0.187	-0.015	0.067	0.021	0.104	0.005	-0.003	0.088
0.09	2	10	0.186	-0.017	0.065	0.019	0.103	0.004	-0.005	0.086
0.09	3	10	0.241	0.050	0.129	0.089	0.164	0.075	0.060	0.145
0.25	1	10	0.137	-0.047	0.034	0.001	0.065	-0.013	-0.037	0.042
0.25	2	10	0.143	-0.039	0.042	0.010	0.072	-0.004	-0.029	0.049
0.25	3	10	0.169	-0.005	0.072	0.040	0.101	0.027	0.004	0.079
0.49	1	10	0.086	-0.058	0.015	0.000	0.034	-0.011	-0.051	0.005
0.49	2	10	0.086	-0.058	0.015	0.000	0.035	-0.011	-0.051	0.005
0.49	3	10	0.074	-0.068	0.001	-0.018	0.022	-0.028	-0.061	-0.004
0.81	1	10	0.028	-0.040	0.000	-0.001	0.006	-0.005	-0.037	-0.015
0.81	2	10	0.029	-0.038	0.003	0.001	0.008	-0.002	-0.035	-0.012
0.81	3	10	0.008	-0.064	-0.023	-0.026	-0.015	-0.030	-0.061	-0.036
0.00	1	20	0.079	0.002	0.022	0.005	0.057	0.005	0.004	0.054
0.00	2	20	0.078	0.000	0.020	0.004	0.056	0.004	0.003	0.053
0.00	3	20	0.080	0.002	0.022	0.006	0.057	0.005	0.005	0.055
0.01	1	20	0.076	-0.001	0.019	0.003	0.054	0.002	0.001	0.051
0.01	2	20	0.077	-0.001	0.020	0.004	0.054	0.003	0.002	0.051
0.01	3	20	0.125	0.048	0.071	0.056	0.103	0.056	0.051	0.098
0.09	1	20	0.068	-0.009	0.015	0.002	0.046	0.001	-0.006	0.040
0.09	2	20	0.066	-0.010	0.014	0.000	0.044	-0.001	-0.008	0.038
0.09	3	20	0.106	0.032	0.056	0.044	0.084	0.042	0.034	0.078
0.25	1	20	0.052	-0.020	0.008	-0.001	0.030	-0.002	-0.018	0.020
0.25	2	20	0.054	-0.018	0.011	0.002	0.033	0.000	-0.016	0.023
0.25	3	20	0.064	-0.006	0.021	0.012	0.043	0.011	-0.004	0.034
0.49	1	20	0.030	-0.030	0.000	-0.004	0.012	-0.006	-0.029	-0.002
0.49	2	20	0.034	-0.025	0.005	0.001	0.016	-0.001	-0.024	0.003
0.49	3	20	0.030	-0.028	0.000	-0.004	0.012	-0.006	-0.027	0.000

Table 10 continued

ρ^2	Shape	n	Corrected r^2							Uncorrected r^2
			Claudy	Ezekiel	Olkin-Pratt	Olkin-Pratt Extended	Olkin-Pratt Pearson	Pratt	Smith	Wherry
0.81	1	20	0.012	-0.017	0.001	0.000	0.003	0.000	-0.016	-0.006
0.81	2	20	0.012	-0.016	0.001	0.001	0.004	0.001	-0.015	-0.005
0.81	3	20	-0.005	-0.035	-0.017	-0.018	-0.014	-0.019	-0.034	-0.023
0.00	1	40	0.032	0.000	0.005	0.000	0.026	0.001	0.000	0.025
0.00	2	40	0.032	0.000	0.005	0.000	0.026	0.001	0.001	0.026
0.00	3	40	0.032	0.000	0.005	0.000	0.026	0.001	0.001	0.026
0.01	1	40	0.031	-0.001	0.004	0.000	0.025	0.000	-0.001	0.024
0.01	2	40	0.033	0.001	0.006	0.002	0.027	0.002	0.001	0.026
0.01	3	40	0.064	0.032	0.038	0.034	0.058	0.034	0.032	0.056
0.09	1	40	0.027	-0.005	0.003	-0.001	0.021	-0.001	-0.004	0.018
0.09	2	40	0.030	-0.002	0.006	0.002	0.023	0.002	-0.002	0.021
0.09	3	40	0.051	0.019	0.027	0.024	0.044	0.024	0.019	0.042
0.25	1	40	0.023	-0.009	0.003	0.001	0.016	0.000	-0.008	0.011
0.25	2	40	0.023	-0.009	0.003	0.001	0.016	0.001	-0.008	0.011
0.25	3	40	0.027	-0.004	0.007	0.004	0.019	0.004	-0.004	0.015
0.49	1	40	0.014	-0.013	0.001	0.000	0.007	0.000	-0.013	0.001
0.49	2	40	0.014	-0.013	0.000	-0.001	0.007	-0.001	-0.013	0.000
0.49	3	40	0.011	-0.015	-0.003	-0.003	0.004	-0.004	-0.015	-0.002
0.81	1	40	0.006	-0.007	0.001	0.001	0.002	0.001	-0.007	-0.002
0.81	2	40	0.005	-0.008	0.000	0.000	0.001	0.000	-0.008	-0.003
0.81	3	40	-0.005	-0.019	-0.011	-0.011	-0.010	-0.011	-0.019	-0.014
0.00	1	60	0.019	-0.001	0.001	-0.001	0.016	0.000	-0.001	0.016
0.00	2	60	0.020	0.000	0.002	0.000	0.017	0.000	0.000	0.017
0.00	3	60	0.020	0.000	0.003	0.001	0.018	0.001	0.001	0.017
0.01	1	60	0.020	0.000	0.002	0.000	0.017	0.000	0.000	0.017
0.01	2	60	0.020	0.000	0.003	0.000	0.017	0.001	0.000	0.017
0.01	3	60	0.042	0.022	0.025	0.024	0.040	0.024	0.023	0.039
0.09	1	60	0.018	-0.002	0.002	0.001	0.015	0.001	-0.002	0.013
0.09	2	60	0.018	-0.003	0.002	0.000	0.014	0.000	-0.002	0.013
0.09	3	60	0.033	0.013	0.018	0.016	0.030	0.016	0.013	0.028
0.25	1	60	0.012	-0.009	-0.001	-0.002	0.008	-0.002	-0.008	0.004
0.25	2	60	0.013	-0.007	0.000	-0.001	0.009	-0.001	-0.007	0.006
0.25	3	60	0.022	0.002	0.009	0.008	0.017	0.008	0.002	0.014

Table 10 continued

ρ^2	Shape	n	Corrected r^2							Uncorrected r^2
			Claudy	Ezekiel	Olkin-Pratt	Olkin-Pratt Extended	Olkin-Pratt Pearson	Pratt	Smith	Wherry
0.49	1	60	0.010	-0.008	0.001	0.000	0.005	0.000	-0.008	0.001
0.49	2	60	0.010	-0.008	0.001	0.000	0.005	0.000	-0.008	0.001
0.49	3	60	0.005	-0.013	-0.004	-0.005	0.000	-0.005	-0.013	-0.004
0.81	1	60	0.003	-0.005	0.000	0.000	0.001	0.000	-0.005	-0.002
0.81	2	60	0.003	-0.006	-0.001	-0.001	0.000	-0.001	-0.006	-0.003
0.81	3	60	-0.005	-0.013	-0.008	-0.008	-0.007	-0.008	-0.013	-0.010
0.00	1	100	0.011	0.000	0.001	0.000	0.010	0.000	0.000	0.010
0.00	2	100	0.011	0.000	0.001	0.000	0.010	0.000	0.000	0.010
0.00	3	100	0.011	0.000	0.001	0.000	0.010	0.000	0.000	0.010
0.01	1	100	0.011	0.000	0.001	0.000	0.010	0.000	0.000	0.010
0.01	2	100	0.011	0.000	0.001	0.000	0.010	0.000	0.000	0.010
0.01	3	100	0.025	0.014	0.015	0.014	0.024	0.014	0.014	0.023
0.09	1	100	0.011	-0.001	0.001	0.001	0.009	0.001	-0.001	0.008
0.09	2	100	0.009	-0.002	0.000	-0.001	0.008	-0.001	-0.002	0.007
0.09	3	100	0.021	0.009	0.012	0.011	0.019	0.011	0.009	0.018
0.25	1	100	0.008	-0.003	0.001	0.000	0.006	0.000	-0.003	0.004
0.25	2	100	0.007	-0.005	-0.001	-0.001	0.005	-0.001	-0.005	0.003
0.25	3	100	0.010	-0.001	0.002	0.002	0.008	0.002	-0.001	0.006
0.49	1	100	0.004	-0.007	-0.002	-0.002	0.001	-0.002	-0.007	-0.002
0.49	2	100	0.004	-0.006	-0.001	-0.001	0.001	-0.001	-0.006	-0.001
0.49	3	100	0.002	-0.008	-0.003	-0.003	-0.001	-0.004	-0.008	-0.003
0.81	1	100	0.002	-0.004	0.000	0.000	0.000	0.000	-0.003	-0.002
0.81	2	100	0.002	-0.003	0.000	0.000	0.000	0.000	-0.003	-0.002
0.81	3	100	-0.002	-0.007	-0.004	-0.004	-0.003	-0.004	-0.007	-0.005
0.00	1	200	0.005	0.000	0.000	0.000	0.005	0.000	0.000	0.005
0.00	2	200	0.005	0.000	0.000	0.000	0.005	0.000	0.000	0.005
0.00	3	200	0.005	0.000	0.000	0.000	0.005	0.000	0.000	0.005
0.01	1	200	0.006	0.000	0.001	0.000	0.005	0.000	0.000	0.005
0.01	2	200	0.005	0.000	0.000	0.000	0.005	0.000	0.000	0.005
0.01	3	200	0.013	0.008	0.008	0.008	0.013	0.008	0.008	0.013
0.09	1	200	0.005	-0.001	0.000	0.000	0.004	0.000	-0.001	0.004
0.09	2	200	0.005	0.000	0.001	0.001	0.005	0.001	0.000	0.004
0.09	3	200	0.010	0.004	0.005	0.005	0.009	0.005	0.004	0.009

Table 10 continued

ρ^2	Shape	n	Corrected r^2						Uncorrected r^2	
			Claudy	Ezekiel	Olkin-Pratt	Olkin-Pratt Extended	Olkin-Pratt Pearson	Pratt	Smith	Wherry
0.25	1	200	0.004	-0.002	0.000	0.000	0.003	0.000	-0.002	0.002
0.25	2	200	0.005	-0.001	0.001	0.001	0.004	0.001	-0.001	0.003
0.25	3	200	0.005	0.000	0.001	0.001	0.004	0.001	0.000	0.003
0.49	1	200	0.002	-0.003	-0.001	-0.001	0.001	-0.001	-0.003	-0.001
0.49	2	200	0.003	-0.002	0.000	0.000	0.001	0.000	-0.002	0.000
0.49	3	200	-0.001	-0.006	-0.003	-0.003	-0.002	-0.003	-0.006	-0.003
0.81	1	200	0.001	-0.001	0.000	0.000	0.000	0.000	-0.001	-0.001
0.81	2	200	0.001	-0.001	0.000	0.000	0.000	0.000	-0.001	0.000
0.81	3	200	-0.002	-0.004	-0.003	-0.003	-0.002	-0.003	-0.004	-0.003

distributed population where $\rho^2 = 0.81$ ($\rho=0.90$), the mean sample r^2 estimate was $\rho^2 = 0.809$ ($\rho=.899$) across the sample of 5,000.

While Table 10 provides mean bias values across the 108 simulation design features, Figure 5 provides a graphical representation of the normally distributed populations grouped by ρ^2 value across sample size conditions. Multiple representations provide for better understanding of study results. At lower ρ^2 values and smaller sample sizes, the majority of the correction formulas show a positive bias. At higher values of ρ^2 and larger sample sizes, all correction formulas perform well.

Table 11
Statistics for Bias of the 540,000 Values for Each of the Eight Estimates Across the 108
(6 X 3 X 6) Simulation Conditions ($n = 5,000/\text{Condition}$)

Statistic/Source	Corrected Estimates							Uncorrected
			Olkin-		Olkin-			
	Claudy	Ezekiel	Pratt	Extended	Pearson	Pratt	Smith	Wherry
<i>Bias</i>								
<i>Mdn</i>	0.019	-0.007	-0.002	-0.004	0.01	-0.005	-0.007	0.008
<i>M</i>	0.044	-0.005	0.014	0.006	0.027	0.003	-0.003	0.022
<i>SD</i>	0.126	0.123	0.124	0.128	0.122	0.128	0.123	0.119
Skewness	1.248	0.227	0.835	0.609	0.968	0.507	0.301	0.778
Kurtosis	4.988	5.564	5.694	5.537	5.69	5.525	5.617	5.686
<i>Absolute Bias</i>								
<i>Mdn</i>	0.048	0.046	0.043	0.047	0.043	0.047	0.045	0.042
<i>M</i>	0.084	0.079	0.078	0.081	0.077	0.081	0.078	0.075
<i>SD</i>	0.104	0.095	0.097	0.099	0.098	0.099	0.094	0.095
Skewness	2.431	2.493	2.557	2.518	2.518	2.504	2.508	2.501
Kurtosis	7.656	8.365	8.556	8.375	8.325	8.325	8.44	8.315
<i>Partial η^2 Values for Simulation Design Factors for Absolute Bias</i>								
Sample n	31.7%	27.7%	24.0%	28.6%	22.9%	29.4%	26.7%	21.5%
Parameter r	7.7%	14.9%	13.7%	14.2%	11.3%	14.4%	14.9%	11.7%
Shape	2.4%	2.8%	2.7%	2.8%	2.5%	2.8%	2.8%	2.7%
$n \times r$	4.8%	2.0%	1.8%	2.3%	1.4%	2.3%	1.9%	1.2%
$n \times \text{Shape}$	0.2%	0.3%	0.3%	0.3%	0.3%	0.3%	0.3%	0.3%
$r \times \text{Shape}$	0.6%	0.7%	0.7%	0.7%	0.6%	0.7%	0.7%	0.7%
$n \times r \times \text{Shape}$	0.2%	0.1%	0.2%	0.1%	0.2%	0.1%	0.1%	0.2%

Note. Wherry is a noncorrection when $p = 1$, as is the case for the bivariate r^2 .

Bias

Table 11 reports the main and interaction effects in the simulation design to predict bias in the corrected and uncorrected estimated r^2 values. Bias was computed by

subtracting the known population ρ^2 value from the sample r^2 estimate. Thus, positive bias values reflect adjusted sample r^2 values that overestimated population parameters. Of course, unbiased estimates by definition should have mean bias values of zero. Descriptive statistics for bias and absolute bias values is also provided in Table 11. Absolute bias provides information that might otherwise be lost when positive and negative bias values cancel each other out.

Precision

Table 12 reports the main and interaction effects in the simulation design to predict precision in the corrected and uncorrected estimated r^2 values. Precision is defined as the standard deviation of the sample r^2 estimate for each of the 5,000 samples across the 108 simulation conditions. Thus, values near zero were expected. Of course, perfectly precise estimates would have mean precision values of zero.

Analysis

In previous studies of statistical bias, researchers have operationally defined unbiased estimators as those estimators which produce values between +0.01 and -0.01 of the corresponding parameter values (Kromney & Hines, 1996; Yin & Fan, 2001). The same criterion was used in the present study to describe an unbiased estimate. Because using this criterion to describe unbiased estimates artificially imposes a categorical restriction on otherwise intervally scaled variables, results were also analyzed by examining the sum of squares. By analyzing results from multiple perspectives you can be more confident that the conclusions drawn from the results are not an artifact of the analytical choices made.

Table 12
Statistics for Precision (SD) for Each of the Eight Estimates Within the 108 (6 X 3 X 6)
Simulation Conditions ($n = 5,000/\text{Condition}$)

Statistic/ Source	Corrected Estimates							Uncorrected
	Claudy	Ezekiel	Olkin- Pratt	Olkin- Pratt Extended	Olkin- Pratt Pearson	Pratt	Smith	Wherry
M	0.092	0.1	0.099	0.103	0.095	0.103	0.099	0.094
SD	0.06	0.071	0.069	0.074	0.065	0.075	0.07	0.063
η^2 Values for Simulation Design Factors for Precision								
Sample n	59.3%	66.6%	65.4%	68.6%	63.7%	68.8%	66.0%	62.4%
Parameter r	30.0%	24.4%	25.1%	22.4%	26.6%	22.3%	24.8%	27.9%
Shape	6.8%	5.6%	5.7%	5.1%	6.0%	5.0%	5.6%	6.3%
$n \times r$	1.8%	1.7%	2.0%	2.2%	1.8%	2.2%	1.7%	1.6%
$n \times \text{Shape}$	0.5%	0.6%	0.4%	0.5%	0.4%	0.5%	0.4%	0.5%
$r \times \text{Shape}$	1.3%	1.1%	1.2%	1.0%	1.3%	1.0%	1.2%	1.4%
$n \times r \times \text{Shape}$	0.3%	0.2%	0.2%	0.2%	0.2%	0.2%	0.2%	0.2%

Note. Wherry is a noncorrection when $p = 1$, as is the case for the bivariate r^2 . Because there is only one case (i.e., the SD of the 5,000 estimates) per cell, the factorial ANOVA degrees of freedom error is zero, and the η^2 values sum to 100%, within rounding error.

As shown in Table 13, Pratt and Olkin-Pratt extended had the greatest number of mean unbiased estimates across the 108 simulation conditions. Pratt produced unbiased estimates 80% of the time, Olkin-Pratt extended produced unbiased estimates 79% of the time. With a comparatively lower percentage of unbiased estimates, Ezekiel, Smith and Olkin-Pratt only generated unbiased estimates 69%, 68% and 69% of the time, respectively.

Table 13
Percentage of *Mean* Unbiased Estimates Returned Within the 108 (6 X 3 X 6)
Simulation Conditions (n = 5,000/Condition)

Correction Formula	ρ						n						Shape ^a			Total
	0	0.1	0.3	0.5	0.7	0.9	10	20	40	60	100	200	1	2	3	
Claudy	17%	11%	22%	33%	50%	78%	6%	6%	17%	33%	56%	94%	33%	36%	36%	35%
Ezekiel	100%	72%	67%	78%	44%	56%	33%	44%	67%	78%	94%	100%	78%	78%	53%	69%
Olkin-Pratt	67%	50%	50%	72%	89%	83%	17%	33%	83%	89%	89%	100%	78%	75%	53%	69%
Olkin- Pratt Extended	83%	61%	61%	89%	94%	83%	33%	78%	83%	89%	89%	100%	92%	92%	53%	79%
Olkin-Pratt Pearson	33%	22%	28%	44%	67%	89%	11%	11%	33%	44%	89%	94%	53%	53%	36%	47%
Pratt	89%	67%	72%	83%	83%	83%	39%	78%	83%	89%	89%	100%	89%	94%	56%	80%
Smith	83%	67%	78%	78%	44%	56%	22%	44%	67%	78%	94%	100%	75%	78%	50%	68%
Wherry (uncorrected)	33%	22%	28%	44%	100%	72%	17%	28%	28%	44%	89%	94%	56%	56%	39%	50%

Note. Unbiased estimates are defined as those estimators that produce values between +0.01 and -0.01 of the corresponding parameter values (Kromney & Hines, 1996; Yin & Fan, 2001).

a. Shape 1: normal, Shape 2: skewness = kurtosis = 1, and Shape 3: skewness = -1.5, kurtosis = 3.5.

Additionally the Olkin-Pratt extended and Pratt formulas were more consistent across sample size conditions, returning unbiased estimates the majority of the time (around 80% or better) at samples sizes of 20 or greater. Ezekiel and Smith were less consistent at lower sample sizes ($n=10, 20, 40$) not returning unbiased estimate percentages around 80% or better unless the sample size was 60 or greater. Generally, at larger sample sizes ($n=200$) all correction formulas produce unbiased estimates greater than 90% of the time regardless of other study design features.

Examination of the correction formulas with the analysis of variance demonstrated that the Olkin-Pratt Pearson formula was least affected by sample size,

partial $\eta^2 = 22.9\%$. The Olkin-Pratt, Smith, Ezekiel, and Olkin-Pratt extended followed with partial η^2 values of 24%, 26.7%, 27.7% and 28.6%, respectively.

Examining the efficacy of the formulas across ρ^2 values, Pratt and Olkin-Pratt extended produced unbiased estimates at least 80% of the time across all ρ^2 conditions except for $\rho^2 = 0.01$ and 0.09 where they produced unbiased estimates around 60% of the time. Ezekiel and Smith did not perform as well. Ezekiel and Smith only produced unbiased estimates greater than 80% of the time at $\rho^2 = 0.0$, dipping below 60% of the time at $\rho^2 = 0.49$ and $\rho^2 = 0.81$. Examination of the correction formulas with the analysis of variance, demonstrated that the Olkin-Pratt Pearson and Claudy had the lowest values of partial η^2 , 7.70% and 11.30%, respectively, across ρ^2 values.

Across shape conditions all formulas produced partial η^2 values less than 3.00%. This small value seen across formulas is consistent with results obtained by Wang and Thompson (2007). Olkin-Pratt Extended and Pratt formulas produced unbiased estimates around 90% of the time for shape conditions 1 (normal) and 2 (skewness = kurtosis = 1). Ezekiel, Smith and Olkin-Pratt performed at about 75% for the same design conditions. The percentage of unbiased estimates under shape = 3 (skewness = 1.5, kurtosis = 3.5) ranged from 36% to 56% for all of the correction formulas.

Conclusions and Recommendations

The results of the present study demonstrate that for large sample sizes ($n=100$ and 200) all correction formulas except for the Claudy formula return unbiased estimates roughly 90% or more of the time. Moreover, at all ρ values except for 0.1 and 0.3 , the Olkin-Pratt extended and the Pratt, produce unbiased estimates greater than 80% of the

time. Similarly, for all shape design conditions except shape = 3 (skewness = -1.5, kurtosis = 3.5), the Olkin- Pratt extended and the Pratt produce unbiased estimates greater than roughly 90% of the time. Thus taking the design conditions into account individually, the Olkin-Pratt and the Pratt formulas appear to produce unbiased estimates a larger proportion of the time than any other correction formula. Recall that unbiased estimates are defined here as those estimators which produce values between +0.01 and - 0.01 of the corresponding parameter values. Yin and Fan (2001) also recommended the Pratt formula in their evaluation of multiple R^2 formulas.

In an evaluation of the cross validation procedure versus using a formula for estimating the predictive power of a regression, Cattin (1980) calculated the Ezekiel (labeled as Wherry) and the Olkin-Pratt extended for the 1 predictor case across 3 sample size conditions (25, 50, 100), and 5 ρ^2 conditions 0.1, 0.2, 0.4, 0.6, and 0.8, to estimate the ρ^2 . Cattin recommended correcting r^2 values of 0.4 and especially for small sample sizes (below $n=50$). Additionally, he recommended that the Olkin-Pratt extended be used instead of the Ezekiel when sample size is small (below 50) to keep the bias below 0.01. Similarly, the findings of the present study point to the use of correction formulas use below $n=60$. Even though Cattin did not use all seven of the formulas that were used in the present study, the Olkin-Pratt Extended was still recommended over the Ezekiel.

It is important to recognize that across design features a formula may work better in some conditions and worse in others. Researchers should evaluate their own design conditions in light of the present studies' design conditions to select the most appropriate

correction formula for their study. Because simple syntax for use in either Excel or SPSS was provided, after selection of the most appropriate formula, corrected r^2 values are easily obtained. As more researchers provide effect size estimates, and interpret them in light of other relevant studies we can move towards a better understanding of our collective research findings.

THE ROBUSTNESS OF ESTIMATES OF PRACTICAL SIGNIFICANCE IN ANOVA

Analysis of variance, ANOVA, was conceived in the early 1920s as Sir Ronald Fisher was developing a way to analyze differences in crop yields (Gamst, Meyers, & Guarino, 2008). ANOVA is a parametric statistical technique that explores mean differences on a single response variable across two or more groups. Social science researchers are often interested in mean differences across groups. For example, a researcher may be interested in the mean difference in retention rates among science and mathematics teachers in rural high schools. Another may want to examine mean daily caloric intake differences among low, middle and high income adults. Reviews of statistical techniques in the literature empirically demonstrate the popularity of ANOVA techniques (Elmore & Woehlke, 1996; Kieffer, Reese, & Thompson, 2001). Furthermore, the future appears to support the continued use of ANOVA techniques as an overwhelming majority of doctoral programs in statistics, measurement and research methodology devote at least a half a semester or a full quarter to teaching analysis of variance (Aiken, West, & Millsap, 2008).

Like all statistical techniques, the validity of ANOVA results is contingent upon the extent to which the assumptions of ANOVA are met. To determine to what extent the ANOVA assumptions are met, a researcher must examine the distributional characteristics of the dependent variable. When this outcome variable exhibits independence, a normal distribution and homogeneity of variance, or equal variance

across groups, the ANOVA assumptions are satisfied. Unfortunately a review of statistical practices revealed that “researchers rarely verify that validity assumptions are satisfied...and...typically use analyses that are nonrobust to assumption violations” (Keselman et al., 1998, p. 350). In practice, the question is not whether ANOVA assumptions are met, but rather, to what extent assumptions are met. An accompanying concern is that not all assumption violations are equivalent. That is, it is well known that the F test is robust to “mild departures from normality” (Harwell, Rubinstein, Hayes, & Olds, 1992, p. 316) and more generally that the F test is considered relatively insensitive to normality assumption violations under equal group size conditions (cf. Glass, Peckham, & Sanders, 1972; Lix, Keselman, & Keselman, 1996). Still researchers have investigated the nonparametric analog to the ANOVA, the Kruskal-Wallis (1952) by ranks procedure to examine its utility when normality violations are of concern (Harwell et al., 1992; Lix et al., 1996).

The behavior of Type I error rates under heterogeneity of variance conditions is well documented in the literature as well (Glass et al., 1972; Harwell et al., 1992). As summarized by Glass et al. (1972) and corroborated by Harwell et al. (1992), when groups are of equal size (balanced design) and heterogeneity of variance is present, there is a slight increase in the Type I error rate. Differential and more pronounced effects are observed when groups are of unequal size (unbalanced design) and heterogeneity of variance is present. In negative pairing, when smaller sample sizes are paired with larger variance, Type I error rates are markedly more than expected. In cases of positive pairing, when smaller sample sizes are paired with smaller variance, Type I error rates

are less than expected. When violations to the homogeneity of variance assumption are a concern, some researchers have recommended the use of alternatives to the ANOVA, such as the James and the Welch tests (Lix et al., 1996).

Keselman and colleagues commented on the severity of violations to ANOVA technique assumptions,

without the assumptions (or barring strong evidence that adequate compensation for them has been made), it can be—and has been— shown that the resulting significance probabilities (p values) are, at best, somewhat different from what they should be and, at worst, worthless (Keselman et al., 1998, p. 351).

Thus violation assumptions should be checked by all researchers and the extent to which violations are present should be matched against empirical literature that details the extent to which Type I error rates and/or power are impacted before making a judgment as to whether a different analytical tool better suited to the distributional characteristics of the data needs to be used.

Confirmatory evidence that ANOVA assumption violations in published literature are common and severe enough to distort power and Type I error rates appreciably is provided in reviews of research practices. For example, in a review of the distributional characteristics of achievement and psychometric measures, researchers observed “*all* [emphasis added] to be significantly nonnormal at the alpha .01 significance level” (Micceri, 1989, p. 156). Similarly violations of the homogeneity of variance assumptions were observed in an empirical review of literature, where on average the highest standard deviation was twice as large as the lowest standard deviation (Keselman

et al., 1998). Perhaps this degree of heterogeneity of variance should not be surprising considering popular texts support this practice declaring the robustness of ANOVA:

recall that the assumption of homogeneity of variance is robust to violation when (1) $F_{\max} \leq 10$, (2) the ratio of largest to smallest sample size is less than 4:1, (3) two-tailed tests are used, and (4) an omnibus analysis is performed” (Tabachnick & Fidell, 2007, p. 123).

The F_{\max} was defined as “ $F_{\max} = s^2_{\text{largest}}/s^2_{\text{smallest}}$ ” (Tabachnick & Fidell, 2007, p.88).

As previously described heterogeneity of variance produces differential effects on the F test based on whether the unequal variances are positively paired or negatively paired with unequal sample sizes. Researchers found that for one-way designs, positive pairings were present roughly a third of the time (31.3%) and negative pairings were present roughly a fifth (22.1%) of the time (Keselman et al., 1998). Still ANOVA is still the most popular (93.3%) inferential analysis technique for between-subjects univariate designs (Keselman et al., 1998).

Moving Beyond Statistical Significance Testing

While ANOVA can be used to test the statistical significance of group mean differences, a second and arguably more important use of ANOVA is to estimate the practical significance, or magnitude of effect, of these group mean differences. Previous researchers have primarily focused on understanding the impact of violation assumptions on test statistics and null hypothesis statistical significance testing, NHSST. This fascination with NHSST is not restricted to one particular discipline. As researchers have commented “null hypothesis significance testing is the workhorse of research in

many disciplines, including medicine, education, ecology, economics, sociology and psychology” (Erceg-Hurn & Mirosevich, 2008, p. 591). Despite its omnipresence, hypothesis testing, typically in the form of NHSST has been a source of contention for many years.

NHSST was introduced in the early 1920s (Fisher, 1925; Fisher & Mackenzie, 1923; Hubbard, Bayarri, Berk, & Carlton, 2003; Neyman & Pearson, 1928). Not long afterward, between 1940 and 1950, a NHSST metastasis evolved (Hubbard & Ryan, 2000). Yet even as NHSST was being rapidly incorporated into researchers’ analytical toolboxes, NHSST was fraught with controversy. Kaufman (1998) stated “the controversy about the use or misuse of statistical significance testing ...has become the major methodological issue of our generation” (p. 1). Books have been published to explain the nature of significance testing (Mohr, 1990) by those who favor (Chow, 1996) and oppose (Harlow, Mulaik, & Steiger, 1997; Kline, 2009; Morrison & Henkel, 1973) its use. Special issues of the *Journal of Experimental Education* (Journal of Experimental Education, 1993) and *Research in the Schools* (Kaufman, 1998) have been dedicated to addressing the NHSST controversy. Even major organizations such as AERA (2006) and APA (2010) have taken a position on the NHSST debate. The most recent declaration given in the newest APA manual recounts,

historically, researchers in psychology have relied heavily on null hypothesis statistical significance testing (NHST) as a starting point for many (but not all) of its analytical approaches. APA stresses that NHST is but a *starting point* [emphasis added] and that additional reporting elements such as *effect sizes*,

[emphasis added] confidence intervals and extensive description are needed to convey the most complete meaning of results (pg. 33).

Further journal editors are also joining the NHSST reform movement as effect size reporting is now required for at least 24 journals (Thompson, 2008).

So it appears as though finally, the call to reform research practices is being heard. Cohen knowingly advised researchers that changing research practice takes time. Reflecting upon the more than four decades that it took the t test to be incorporated into textbooks, Cohen offered the following words of comfort to discouraged researchers, “if you publish something that you think is really good, and a year or a decade or two go by and hardly anyone seems to have taken notice, remember the t test, and take heart” (Cohen, 1990, p. 1311).

Previous Studies

While the focus is presently shifting more towards effect sizes, the historical preoccupation with statistical significance testing has focused previous research efforts on evaluating the impact of assumption violations to measures related to the F test in ANOVA. Indeed there is limited research available on the robustness of estimates of practical significance. One such article was published over thirty years ago by Carroll and Nordholm (1975). Means across non-null conditions were held constant, while the within population variances were adjusted to achieve resulting values of $\eta^2 = 0.05, 0.15, 0.40$ and 0.75 within the context of a three level one-way fixed effects ANOVA under both balanced and unbalanced conditions. Carroll and Nordholm (1975) found that just as heterogeneity of variance in unbalanced designs is a cause for serious distortions to

power and Type I error rates in ANOVA, similarly, the most serious distortions to ε^2 and ω^2 occurred when variances were unequal across unbalanced designs.

Keselman (1975) also investigated the robustness of effect sizes in ANOVA by examining bias and precision of η^2 , ε^2 and ω^2 for normal and exponential populations. Keselman's design conditions include (a) one group configuration (b) three different effect sizes conditions (small, medium and large) and (c) two types of mean variability conditions, intermediate and maximum (cf. Cohen, 1988). Keselman found that "omega squared is a more accurate estimator of the true population magnitude while eta squared has the smallest sampling variability" (p. 47).

Wilcox has published extensively on the robustness (or lack thereof) of the F test under assumption violations (Wilcox, 1995; Wilcox, Charlin, & Thompson, 1986) and suggested the use of more robust methods (Wilcox, 1993; Wilcox & Keselman, 2003). Wilcox has also examined the robustness of one measure of effect size, Cohen's d . In cases where there is a contaminated normal distribution (see Tukey, 1960), "Cohen's d can mask a large effect size" (Wilcox, 2006, p. 355). Wilcox has shown that when the tails of the distribution are thicker, as in a contaminated normal distribution, or in the presence of outliers, indices of effect size, such as Cohen's d , can be distorted (Wilcox, 2006).

The purpose of the present article is to move beyond the robustness of estimates of statistical significance (Type I and power) to evaluate the robustness of estimates of practical significance (η^2 , ε^2 , and ω^2) in one-way between subjects univariate ANOVA. Theoretically expected estimates of η^2 , ε^2 , and ω^2 when assumptions are perfectly met are

compared to actual empirical estimates when assumptions are not met to understand the utility of effect size measures in the presence of assumption violations.

Methods

To the extent possible, the conditions for the present Monte Carlo investigation were chosen based on previous research findings that either demonstrated a need to investigate a particular condition or to investigate specified researcher practices derived from the literature. Thus the conditions modeled are based on what previous research indicates should have an impact while still maintaining an ecologically valid footing through grounding in actual observed researcher practices.

Formulas for Computing Effect Sizes

While the design of this study allows for confirmation of previous findings regarding the behavior of estimates of statistical significance (i.e., Type I and power) under ANOVA assumption violations the focus is on the behavior of estimates of practical significance (η^2 , ε^2 and ω^2). Eta squared, η^2 , measures the “proportion of the variance in the population that is accounted for by variation in the treatment” (Grissom & Kim, 2005, p. 121). Eta squared is given by $\eta^2 = SS_{\text{model}} / (SS_{\text{total}})$ or in the case of one-way design, can also be computed as $((k - 1) * (F)) / (((k - 1) * (F)) + N - k)$ where k is the number of groups and N is the total sample size (Wilcox, 1987). It is well known that η^2 like R^2 (Yin & Fan, 2001), and r^2 (Wang & Thompson, 2007), is positively biased. To correct this bias Kelley (1935) and Hays (1981) developed ε^2 and ω^2 , respectively. Epsilon squared is given by $\varepsilon^2 = (SS_{\text{model}} - (k - 1) * (MS_{\text{err}})) / (SS_{\text{total}})$ or equivalently by $(F - 1) / (F + ((N - k) / (k - 1)))$ (Carroll & Nordholm, 1975). Omega squared is given by

$\omega^2 = (SS_{\text{model}} - (k - 1) * (MS_{\text{err}})) / (SS_{\text{total}} + MS_{\text{err}})$ or equivalently by $(F - 1) / (F + ((N - k + 1) / (k - 1)))$ (Carroll & Nordholm, 1975). The presence of the F test statistic in the formulas should underscore the inherent relationship between all parametric analyses under the general linear model as well as the need to understand the behavior of the given effect sizes under assumption violations.

Means

Although Cohen (1988) himself eschewed the thoughtless fixation on benchmark numbers, Cohen's d values of 0.2, 0.5 and 0.8, for better or for worse, have prevailed as benchmarks for small, medium, and large effect sizes. In the present study all distributions had a mean of 100 for the null condition. For the non-null conditions, chosen Cohen's d values of 0.2, 0.5, 0.8 and 1.0, were converted to Cohen's f effect size index, "the standard deviation of the standardized k population means" using tabled conversion values that accounted for the number of group means (Cohen, 1988, p. 276). In determining the necessary mean differences to obtain the given Cohen's f value, case 1, as designated by Cohen (1988) was used where one mean is at each end of the range and "the remaining $k - 2$ means are all at the midpoint" (p. 277).

Number of Groups

Wilcox, Charlin and Thompson (1986) demonstrated that when there were four groups the F test was not as robust as when there were two groups. Therefore, the condition of number of groups is an important one to consider when evaluating the robustness of the F test. In a review of Monte Carlo studies, the number of groups examined by researcher varied between 2 and 10 for equal group sizes and between 2

and 6 for unequal group sizes with 3 groups being the most commonly examined (Harwell et al., 1992). The number of groups examined in the present study are 2, 3 and 4.

Group Size Proportions

In a review of the analytical practices of educational researchers, Keselman and his colleagues (1998) explained that one-way designs made up 58.3% of the 61 between subjects univariate studies examined. Furthermore, in the 23 one-way studies with an unbalanced design, the largest group size was more than three times larger than the smallest group size in 43.5% of the studies. For $k=2$, I selected group sizes of 12:12, 8:16, 6:18 and 4:20 for a total N of 24 and 24:24, 16:32, 12:36 and 8:40 for a total N of 48. For $k=3$, I selected group sizes of 12:12:12, 9:9:18, 6:12:18 and 6:6:24 for a total N of 36 and 24:24:24, 18:18:36, 12:24:36 and 12:12:48 for a total N of 72. For $k=4$, I selected group sizes of 12:12:12:12, 8:8:16:16, 8:8:8:24 and 6:12:12:18 for a total N of 48 and of 24:24:24:24, 16:16:32:32, 16:16:16:48 and 12:24:24:36 for a total N of 96.

Sample Size

Earlier studies on the robustness of ANOVA to assumption violations had relatively small total N s. For example, Hsu's (1938) largest total sample size was 20 and Box's (1954) largest total sample size was 25. A later study, that was highlighted as "exemplary" in Glass, Peckham and Sander's (1972, p. 265) highly cited article, examined a minimum sample size of 8 for the two group condition and a maximum total sample size of 128 for the four group condition (Donaldson, 1968).

In a meta-analytic summary of 28 Monte Carlo studies, researchers reported the average total sample size to be 111, *SD* 154, across the simulation studies reviewed, with a minimum total sample size of 8, and a maximum of 750 (Harwell, Rubinstein, Hayes, & Olds, 1992). In practice, a summary of 10 years of analytical techniques used in *AERJ* and *JCP* resulted in a median total sample size of 108. 5, *SD* 35.8, across a median number of variables of 4, *SD* 1.5 (Kieffer, Reese, & Thompson, 2001). The minimum total sample size used for the present study was 24, for the two group condition. The maximum total sample size used in the present study was 96 for the four group condition. Because earlier studies used smaller *ns* per group, a small group cell size of 12 was chosen to compare to the larger group cell size of 24.

Variance

Heterogeneity of variance is a serious assumption violation in the analysis of variance (Harwell et al., 1992; Lix et al., 1996). Yet, for one way designs, standard deviations twice as large as the smallest standard deviation are the average (Keselman et al., 1998). Furthermore, it is well documented that not only the heterogeneity of variances is an important condition but that the way in which sample sizes are paired with variances produces differential results. Smaller sample sizes paired with larger variances (negative pairing) produce larger Type I error rates; while smaller sample sizes paired with smaller variances (positive pairing) produce a lower Type I error rate (Glass et al., 1972; Harwell et al., 1992). Therefore both negative and positive pairings of heterogeneity are modeled in the present study. Equal variance conditions were $\sigma^2=225$, for $k = 2, 3$ and 4. Unequal negative pairing conditions were $\sigma^2=900:225; 900:450:225$;

and 900:450:450:225, for $k = 2, 3$ and 4, respectively. Positive pairing conditions were $\sigma^2=225: 900; 225:450: 900;$ and $225:450:450: 900$, for $k = 2, 3$ and 4, respectively.

Shape

As previously mentioned mild departures from normality had negligible effects on the F test (Harwell et al., 1992), thus conditions in the present study were chosen to be normal, mildly deviant (skewness = kurtosis = 0.5) and moderately deviant (skewness = 1, kurtosis = 3.75). SAS generated population data with the given shape parameters using Vale and Maurelli's (1983) multivariate extension of Fleishman's procedure (1978). To confirm that the program was working as programmed, population of 100,000 scores were generated to compare the obtained shape parameters with the specified shape parameters.

Replications

In order to obtain stable estimates of Type 1 error rates, 5,000 replications per condition were generated (Robey & Barcikowski, 1992). Detailed SAS programming explanations, including sample programs can be found in *SAS for Monte Carlo Studies* (Fan, Felsovalyi, Sivo, & Keenan, 2001).

Summary Conditions by Group Size

For group size, $k=2$, using the RANNOR random number generator in SAS (version 9.2), 5,000 samples were drawn from each of the 360 ($5 \times 4 \times 3 \times 3 \times 2$) simulation conditions (i.e., population Cohen's d values of 0.0, 0.2, 0.5, 0.8 and 1.0; group proportion ratios of 1:1, 1:2, 1:3, and 1:5; and population shapes normal, skewness

= kurtosis = 0.5, and skewness = 1 with kurtosis = 3.75; variance ratios of 1:1, 4:1, 1:4 and total N values of 24 and 48).

For group size, $k=3$, using the RANNOR random number generator in SAS (version 9.2), 5,000 samples were drawn from each of the 360 ($5 \times 4 \times 3 \times 3 \times 2$) simulation conditions (i.e., population d values of 0.0, 0.2, 0.5, 0.8 and 1.0; group proportion ratios of 1:1:1, 1:1:2, 1:2:3, and 1:1:4; and population shapes normal, skewness = kurtosis = 0.5, and skewness = 1 with kurtosis = 3.75; variance ratios of 1:1:1, 4:2:1, 1:2:4 and total N values of 36 and 72).

For group size, $k=4$, using the RANNOR random number generator in SAS (version 9.2), 5,000 samples were drawn from each of the 360 ($5 \times 4 \times 3 \times 3 \times 2$) simulation conditions (i.e., population d values of 0.0, 0.2, 0.5, 0.8 and 1.0; group proportion ratios of 1:1:1:1, 1:1:2:2, 1:1:1:3, and 1:2:2:3; and population shapes normal, skewness = kurtosis = 0.5, and skewness = 1 with kurtosis = 3.75; variance ratios of 1:1:1:1, 4:2:2:1, 1:2:2:4 and total N values of 48 and 96).

Thus, in the present study three indices of practical significance (η^2 , ε^2 , ω^2) and two indices of statistical significance (Type I error and power) were computed for each of the 5,400, 000 ($5,000 \times 360 \times 3$) samples. Estimated (a) bias (b) absolute bias and (c) precision was also computed.

Simulation Baseline Check

When ANOVA assumptions are met, the expectation is that nominal α levels agree with actual obtained Type I error rates. Similarly, when ANOVA assumptions are met, the expectation is that theoretical power levels agree with actual obtained power.

When conducting a Monte Carlo study, providing both Type I error rates and theoretical versus empirical power estimates presents evidence that the simulation study was correctly conducted. Glass et al. recommended that such “‘baseline checks’ of the entire simulation procedure should be performed and reported” (1972, p. 282).

Table 14
Empirical Type I Error Rates

k	Population Cohen's d	Group Size Proportion	Variance Ratio	Shape	Type I Error Rate	
					Smaller N	Larger N
2	0	1:1	1:1	Normal	0.047	0.051
2	0	1:2	1:1	Normal	0.053	0.047
2	0	1:3	1:1	Normal	0.050	0.052
2	0	1:5	1:1	Normal	0.052	0.053
3	0	1:1:1	1:1:1	Normal	0.050	0.050
3	0	1:1:2	1:1:1	Normal	0.053	0.049
3	0	1:1:4	1:1:1	Normal	0.051	0.049
3	0	1:2:3	1:1:1	Normal	0.050	0.049
4	0	1:1:1:1	1:1:1:1	Normal	0.049	0.055
4	0	1:1:1:3	1:1:1:1	Normal	0.050	0.047
4	0	1:1:2:2	1:1:1:1	Normal	0.048	0.043
4	0	1:2:2:3	1:1:1:1	Normal	0.050	0.049

Table 14 provides empirically obtained Type I error rates for the various group configurations and various group size proportion combinations, under equal variance and normality conditions. As you can see, empirical estimates are, as expected, close to or equal to 0.05 when nominal $\alpha=0.05$.

Table 15
Empirical Power Estimates (Normal Distribution and Equal Variances)

k	n	d	ϕ	Empirical Power	Theoretical ^a Power
2	12 ₍₁₆₎	0.2	0.35 _(0.50)	0.078 _(0.102)	0.076
2	12 ₍₁₆₎	0.5	0.87 _(1.00)	0.222 _(0.276)	0.216
2	12 ₍₁₆₎	0.8	1.39 _(1.50)	0.462 _(0.544)	0.466
2	12	1.0	1.73	0.657	0.649
2	24 ₍₃₂₎	0.2	0.49 _(0.50)	0.107 _(0.108)	0.104
2	24 ₍₃₂₎	0.5	1.22 _(1.50)	0.387 _(0.556)	0.396
2	24 ₍₃₂₎	0.8	1.96 _(2.00)	0.771 _(0.795)	0.774
2	24	1.0	2.45	0.922	0.924
3	12	0.2	0.02	0.065	0.067
3	12	0.5	0.14	0.167	0.167
3	12	0.8	0.33	0.375	0.369
3	12	1.0	0.49	0.552	0.542
3	24	0.2	0.03	0.080	0.086
3	24	0.5	0.20	0.314	0.310
3	24	0.8	0.47	0.683	0.678
3	24	1.0	0.70	0.867	0.868
4	12 ₍₁₆₎	0.2	0.25 _(0.56)	0.064 _(0.129)	0.063
4	12	0.5	0.61	0.141	0.144
4	12 ₍₁₆₎	0.8	0.98 _(1.12)	0.309 _(0.416)	0.319
4	12	1.0	1.23	0.481	0.481
4	24 ₍₃₂₎	0.2	0.35 _(0.56)	0.075 _(0.131)	0.079
4	24 ₍₃₂₎	0.5	0.87 _(1.12)	0.261 _(0.434)	0.265
4	24 ₍₃₂₎	0.8	1.39 _(1.68)	0.621 _(0.804)	0.615
4	24	1.0	1.73	0.828	0.824

Note. Values in parentheses are the design conditions and empirical power estimates from Donaldson (1968) that most closely match the present study's design conditions.

a. Values obtained using G* Power.

In the widely cited article by Glass and colleagues (1972), Donaldson's (1968) paper is credited as being "exemplary of many of the best features of a simulation robustness study" (p. 265). Donaldson provided tabled values of empirical power estimates. To the extent possible, the closest design conditions from Donaldson's study are given in Table 15 to provide a comparison to the empirical power values in the present study when ANOVA assumptions are perfectly met. Also provided are theoretical power estimates obtained using G*Power 3 (Version 3.1.0; Faul, Erdfelder, Lang, & Buchner, 2007).

As a final confirmation that the simulation worked as it was intended, study results for the most severe case of assumption violations are presented, when unequal samples sizes are paired with heterogeneous variances. Table 16 displays these values for both the smaller total sample size and larger total sample size condition across $k = 2$, 3 and 4.

Analyses

Estimated parameter biases were computed for each of the 5,400,000 analyses modeled, by subtracting the individual sample η^2 , ε^2 , ω^2 from the expected parameter η^2 values. Thus positive parameter bias values indicated sample estimates that underestimated the parameter value and negative parameter bias values indicated sample estimates that overestimated parameter values. Additionally, absolute parameter bias was computed by taking the absolute value of each estimated parameter bias value. Finally, precision was estimated by evaluating the standard deviations of each of the 360 conditions modeled for $k = 2$, 3, and 4.

Table 16
Impact of Heterogeneity of Variance on Type I Error Rates

k	Cohen's d	Group Size Proportion	Variance Ratio	Type I Error	
				Smaller N	Larger N
2	0	1:1	1:4	0.0510	0.0464
2	0	1:2	1:4	0.0204	0.0144
2	0	1:3	1:4	0.0092	0.0086
2	0	1:5	1:4	0.0032	0.0030
2	0	1:1	4:1	0.0582	0.0490
2	0	1:2	4:1	0.1126	0.1106
2	0	1:3	4:1	0.1548	0.1532
2	0	1:5	4:1	0.2040	0.2154
3	0	1:1:1	1:2:4	0.0562	0.0584
3	0	1:1:2	1:2:4	0.0292	0.0272
3	0	1:1:4	1:2:4	0.0118	0.0114
3	0	1:2:3	1:2:4	0.0200	0.0274
3	0	1:1:1	4:2:1	0.0596	0.0558
3	0	1:1:2	4:2:1	0.0980	0.0946
3	0	1:1:4	4:2:1	0.1628	0.1536
3	0	1:2:3	4:2:1	0.1346	0.1198
4	0	1:1:1:1	1:2:2:4	0.061	0.053
4	0	1:1:1:3	1:2:2:4	0.016	0.022
4	0	1:1:2:2	1:2:2:4	0.031	0.039
4	0	1:2:2:3	1:2:2:4	0.029	0.032
4	0	1:1:1:1	4:2:2:1	0.055	0.062
4	0	1:1:1:3	4:2:2:1	0.120	0.117
4	0	1:1:2:2	4:2:2:1	0.082	0.088
4	0	1:2:2:3	4:2:2:1	0.108	0.110

Results

Parameter Bias

Five-way full factorial ANOVAs were conducted for each of the three group types using all 1,800,000 replications per k across the simulation's $5 \times 4 \times 3 \times 3 \times 2 = 360$ design conditions. Table 17 presents the total sum of squares (SS) and the η^2 's obtained for the parameter biases for the three effect sizes (i.e., η^2 , ε^2 and ω^2) when $k = 2$. The single most important predictor of parameter bias across η^2 , ε^2 and ω^2 was Cohen's d , which accounted for 16%, 14% and 15%, respectively, of the total variance observed per effect size. Variance as a design condition accounted for 5% of the total variance observed across all effect sizes (i.e., η^2 , ε^2 and ω^2). The remainder of the design conditions and their interactions accounted for less than 5% of the total variance observed across all effect sizes.

ANOVA summary Table 18 details that for $k=3$ condition, Cohen's d was not responsible for as much of the total variance observed in the parameter bias as it had been in the two group condition. However, variance as a design condition is still accounting for 5% of the total variance per effect size as in the $k=2$ condition. The remainder of the design conditions and their interactions accounted for less than 5% of the total variance observed across all effect sizes.

Table 17
Estimated Parameter Bias in Three ANOVA Effect Sizes for Two Groups

Source	df	η^2		ε^2		ω^2	
		SS	η^2	SS	η^2	SS	η^2
Parameter Cohen's d	4	2511.5	0.16	2337.3	0.14	2467.5	0.15
Variance/Heterogeneity	2	735.4	0.05	787.0	0.05	748.7	0.05
Group Size Ratios	3	49.2	0.00	52.4	0.00	49.7	0.00
Population Shape	2	2.2	0.00	2.3	0.00	2.2	0.00
Total N	1	240.5	0.01	1.8	0.00	0.9	0.00
Cohen's d *Variance	8	459.5	0.03	490.8	0.03	469.2	0.03
Cohen's d *Group Prop	12	89.3	0.01	95.4	0.01	91.0	0.01
Cohen's d *Shape	8	2.2	0.00	2.3	0.00	2.2	0.00
Cohen's d *Total N	4	0.5	0.00	0.0	0.00	0.3	0.00
Variance*Group Prop	6	243.3	0.02	261.2	0.02	248.7	0.02
Variance*Shape	4	5.0	0.00	5.4	0.00	5.1	0.00
Variance*Total N	2	18.1	0.00	20.8	0.00	19.0	0.00
Group Prop*Shape	6	0.7	0.00	0.8	0.00	0.8	0.00
Group Prop*Total N	3	0.6	0.00	0.5	0.00	0.6	0.00
Shape*Total N	2	0.1	0.00	0.1	0.00	0.1	0.00
Cohen's d *Variance*Group Prop	24	22.2	0.00	23.7	0.00	23.1	0.00
Cohen's d *Variance*Shape	16	1.8	0.00	2.0	0.00	1.9	0.00
Cohen's d *Variance*Total N	8	0.9	0.00	1.0	0.00	0.9	0.00
Cohen's d * Group Prop*Shape	24	0.2	0.00	0.3	0.00	0.2	0.00
Cohen's d * Group Prop*Total N	12	0.1	0.00	0.1	0.00	0.1	0.00
Cohen's d * Shape *Total N	8	0.2	0.00	0.2	0.00	0.2	0.00
Variance* Group Prop *Shape	12	0.2	0.00	0.2	0.00	0.2	0.00
Variance* Group Prop *Total N	6	11.4	0.00	13.4	0.00	12.0	0.00
Variance* Shape*Total N	4	0.5	0.00	0.5	0.00	0.5	0.00
Group Prop* Shape*Total N	6	0.0	0.00	0.1	0.00	0.1	0.00
Cohen's d * Variance* Group Prop*Shape	48	0.3	0.00	0.3	0.00	0.3	0.00
Cohen's d * Variance* Group Prop*Total N	24	0.2	0.00	0.2	0.00	0.2	0.00
Cohen's d *Variance*Shape*Total N	16	0.2	0.00	0.2	0.00	0.2	0.00
Cohen's d *Group Prop*Shape*Total N	24	0.1	0.00	0.1	0.00	0.1	0.00
Variance* Group Prop*Shape*Total N	12	0.1	0.00	0.1	0.00	0.1	0.00
Cohen's d * Variance* Group Prop*Shape*Total N	48	0.2	0.00	0.2	0.00	0.2	0.00
Error	1799640	11672.4	0.73	12585.5	0.75	11959.0	0.74
Total	1799999	16068.9		16686.0		16105.1	

Table 18
Estimated Bias in Three ANOVA Effect Sizes for Three Groups

Source	df	η^2		ε^2		ω^2	
		SS	η^2	SS	η^2	SS	η^2
Parameter Cohen's d	4	921.8	0.09	809.2	0.08	854.0	0.09
Variance/ Heterogeneity	2	461.6	0.05	504.9	0.05	487.2	0.05
Group Size Ratios	3	5.9	0.00	6.4	0.00	6.2	0.00
Population Shape	2	0.9	0.00	1.0	0.00	1.0	0.00
Total N	1	372.9	0.04	0.4	0.00	0.1	0.00
Cohen's d *Variance	8	288.4	0.03	314.7	0.03	304.7	0.03
Cohen's d *Group Prop	12	11.3	0.00	12.4	0.00	12.0	0.00
Cohen's d *Shape	8	0.7	0.00	0.8	0.00	0.7	0.00
Cohen's d *Total N	4	0.5	0.00	0.0	0.00	0.1	0.00
Variance*Group Prop	6	110.3	0.01	121.3	0.01	116.8	0.01
Variance*Shape	4	1.6	0.00	1.8	0.00	1.8	0.00
Variance*Total N	2	10.9	0.00	13.1	0.00	12.2	0.00
Group Prop*Shape	6	0.1	0.00	0.1	0.00	0.1	0.00
Group Prop*Total N	3	0.1	0.00	0.1	0.00	0.1	0.00
Shape*Total N	2	0.0	0.00	0.1	0.00	0.1	0.00
Cohen's d *Variance*Group Prop	24	7.1	0.00	7.7	0.00	7.5	0.00
Cohen's d *Variance*Shape	16	0.7	0.00	0.8	0.00	0.7	0.00
Cohen's d *Variance*Total N	8	0.3	0.00	0.3	0.00	0.3	0.00
Cohen's d * Group Prop*Shape	24	0.2	0.00	0.2	0.00	0.2	0.00
Cohen's d * Group Prop*Total N	12	0.0	0.00	0.0	0.00	0.0	0.00
Cohen's d * Shape*Total N	8	0.1	0.00	0.1	0.00	0.1	0.00
Variance* Group Prop*Shape	12	0.0	0.00	0.0	0.00	0.0	0.00
Variance* Group Prop*Total N	6	6.1	0.00	7.6	0.00	7.0	0.00
Variance*Shape*Total N	4	0.2	0.00	0.2	0.00	0.2	0.00
Group Prop* Shape*Total N	6	0.0	0.00	0.0	0.00	0.0	0.00
Cohen's d *Variance* Group Prop*Shape	48	0.1	0.00	0.1	0.00	0.1	0.00
Cohen's d *Variance* Group Prop*Total N	24	0.1	0.00	0.1	0.00	0.1	0.00
Cohen's d *Variance * Shape * Total N	16	0.1	0.00	0.1	0.00	0.1	0.00
Cohen's d *Group Prop*Shape*Total N	24	0.1	0.00	0.2	0.00	0.1	0.00
Variance*Group Prop*Shape*Total N	12	0.0	0.00	0.0	0.00	0.0	0.00
Cohen's d *Variance * Group Prop* Shape * Total N	48	0.1	0.00	0.2	0.00	0.2	0.00
Error	1799640	7611.0	0.78	8419.0	0.82	8104.0	0.82
Total	1799999	9814.0		10220.0		9918.0	

Table 19
Estimated Bias in Three ANOVA Effect Sizes for Four Groups

Source	df	η^2		ϵ^2		ω^2	
		SS	η^2	SS	η^2	SS	η^2
Parameter Cohen's d	4	492.3	0.07	417.5	0.06	438.1	0.06
Variance/Heterogeneity	2	274.8	0.04	303.8	0.04	295.5	0.04
Group Size Ratios	3	6.4	0.00	7.1	0.00	6.9	0.00
Population Shape	2	0.4	0.00	0.5	0.00	0.5	0.00
Total N	1	452.0	0.07	0.1	0.00	0.0	0.00
Cohen's d *Variance	8	180.5	0.03	199.1	0.03	194.1	0.03
Cohen's d * Group Prop	12	7.1	0.00	7.8	0.00	7.6	0.00
Cohen's d *Shape	8	0.4	0.00	0.4	0.00	0.4	0.00
Cohen's d *Total N	4	0.3	0.00	0.0	0.00	0.0	0.00
Variance * Group Prop	6	50.0	0.01	55.6	0.01	54.0	0.01
Variance * Shape	4	0.9	0.00	1.0	0.00	1.0	0.00
Variance * Total N	2	5.7	0.00	7.1	0.00	6.7	0.00
Group Prop*Shape	6	0.0	0.00	0.0	0.00	0.0	0.00
Group Prop*Total N	3	0.0	0.00	0.0	0.00	0.0	0.00
Shape * Total N	2	0.1	0.00	0.1	0.00	0.1	0.00
Cohen's d *Variance *Group Prop	24	3.6	0.00	3.9	0.00	3.8	0.00
Cohen's d *Variance * Shape	16	0.4	0.00	0.4	0.00	0.4	0.00
Cohen's d *Variance * Total N	8	0.2	0.00	0.1	0.00	0.1	0.00
Cohen's d * Group Prop*Shape	24	0.1	0.00	0.1	0.00	0.1	0.00
Cohen's d * Group Prop*Total N	12	0.0	0.00	0.0	0.00	0.0	0.00
Cohen's d *Shape * Total N	8	0.0	0.00	0.0	0.00	0.0	0.00
Variance * Group Prop*Shape	12	0.0	0.00	0.1	0.00	0.1	0.00
Variance* Group Prop * Total N	6	2.8	0.00	3.5	0.00	3.3	0.00
Variance * Shape * Total N	4	0.1	0.00	0.1	0.00	0.1	0.00
Group Prop*Shape * Total N	6	0.0	0.00	0.0	0.00	0.0	0.00
Cohen's d *Variance * Group Prop*Shape	48	0.1	0.00	0.1	0.00	0.1	0.00
Cohen's d *Variance *Group Prop * Total N	24	0.1	0.00	0.1	0.00	0.1	0.00
Cohen's d *Variance * Shape * Total N	16	0.1	0.00	0.1	0.00	0.1	0.00
Cohen's d * Group Prop*Shape * Total N	24	0.1	0.00	0.1	0.00	0.1	0.00
Variance * Group Prop*Shape * Total N	12	0.0	0.00	0.0	0.00	0.0	0.00
Cohen's d * Variance * Group Prop*Shape * Total N	48	0.1	0.00	0.1	0.00	0.1	0.00
Error	1799640	5450.9	0.79	6107.3	0.86	5924.0	0.85
Total	1799999	6929.5		7116.1		6937.1	

For $k = 4$, Cohen's d played an even smaller role in the total amount of variance present in the parameter bias, by accounting for only 7%, 6% and 6% of the total

variance across η^2 , ε^2 and ω^2 , respectively as shown in Table 19. Also noteworthy in the $k = 4$ condition, the total N design condition surfaced as a relatively demonstrable percentage of the explained variance present in the bias for η^2 only. Variance as a design condition dropped below 5% of the total variance explained across all effect sizes.

Absolute Parameter Bias

While parameter bias is an important feature to examine, a smaller parameter bias value is observed when positive and negative bias estimates cancel each other out. A solution to this problem is to calculate the absolute parameter bias. In doing so, all deviations from the parameter, whether positive or negative are accounted for. Five-way full factorial ANOVAs were conducted for each of the three group types using all 1,800,000 replications per group type across the simulation's $5 \times 4 \times 3 \times 3 \times 2 = 360$ design conditions. Table 20 displays the total sum of squares (SS) and the η^2 's obtained for the absolute parameter biases for the three effect sizes (i.e., η^2 , ε^2 and ω^2) when $k = 2$. In comparison to the parameter bias summary ANOVA table given in Table 17, the amount of variance explained in the absolute parameter bias by Cohen's d for η^2 has gone up by more than 5% and more than doubled for ε^2 and ω^2 to 30% and 32%, respectively. Thus while in Table 17, for ε^2 and ω^2 , Cohen's d appeared to exhibit less parameter bias than η^2 , in fact the presence of both positive and negative parameter biases in ε^2 and ω^2 is evidenced in the absolute parameter bias results. The remainder of the design conditions and their interactions accounted for less than 5% of the total variance observed across all effect sizes.

Table 20
Estimated Absolute Bias in Three ANOVA Effect Sizes for Two Groups

Source	df	η^2		ε^2		ω^2	
		SS	η^2	SS	η^2	SS	η^2
Parameter Cohen's d	4	1749.3	0.22	2501.1	0.30	2548.4	0.32
Variance/ Heterogeneity	2	65.3	0.01	77.6	0.01	80.6	0.01
Group Size Ratios	3	15.0	0.00	35.2	0.00	35.4	0.00
Population Shape	2	0.1	0.00	0.0	0.00	0.0	0.00
Total N	1	169.6	0.02	280.5	0.03	251.5	0.03
Cohen's d *Variance	8	111.9	0.01	131.9	0.02	136.6	0.02
Cohen's d *Group Prop	12	8.0	0.00	13.1	0.00	13.7	0.00
Cohen's d *Shape	8	0.1	0.00	0.0	0.00	0.0	0.00
Cohen's d *Total N	4	5.2	0.00	0.2	0.00	0.3	0.00
Variance*Group Prop	6	46.6	0.01	10.8	0.00	8.9	0.00
Variance*Shape	4	2.0	0.00	1.3	0.00	1.2	0.00
Variance*Total N	2	24.8	0.00	8.0	0.00	7.2	0.00
Group Prop*Shape	6	0.2	0.00	0.1	0.00	0.1	0.00
Group Prop*Total N	3	0.2	0.00	0.0	0.00	0.0	0.00
Shape*Total N	2	0.1	0.00	0.0	0.00	0.0	0.00
Cohen's d *Variance*Group Prop	24	31.7	0.00	25.2	0.00	25.6	0.00
Cohen's d *Variance*Shape	16	1.0	0.00	0.8	0.00	0.8	0.00
Cohen's d *Variance*Total N	8	9.7	0.00	4.8	0.00	4.5	0.00
Cohen's d * Group Prop*Shape	24	0.1	0.00	0.1	0.00	0.1	0.00
Cohen's d * Group Prop*Total N	12	1.5	0.00	0.8	0.00	0.7	0.00
Cohen's d * Shape *Total N	8	0.1	0.00	0.0	0.00	0.0	0.00
Variance* Group Prop *Shape	12	0.1	0.00	0.1	0.00	0.1	0.00
Variance* Group Prop *Total N	6	11.6	0.00	4.1	0.00	3.6	0.00
Variance* Shape*Total N	4	0.3	0.00	0.1	0.00	0.1	0.00
Group Prop* Shape*Total N	6	0.0	0.00	0.0	0.00	0.0	0.00
Cohen's d * Variance* Group Prop*Shape	48	0.4	0.00	0.3	0.00	0.3	0.00
Cohen's d * Variance* Group Prop*Total N	24	0.2	0.00	0.2	0.00	0.2	0.00
Cohen's d *Variance*Shape*Total N	16	0.1	0.00	0.0	0.00	0.0	0.00
Cohen's d *Group Prop*Shape*Total N	24	0.0	0.00	0.0	0.00	0.0	0.00
Variance* Group Prop*Shape*Total N	12	0.0	0.00	0.0	0.00	0.0	0.00
Cohen's d * Variance* Group Prop*Shape*Total N	48	0.1	0.00	0.1	0.00	0.1	0.00
Error	1799640	5871.3	0.72	5115.6	0.62	4898.1	0.61
Total	1799999	8126.7		8212.3		8018.5	

Table 21
Estimated Absolute Bias in Three ANOVA Effect Sizes for Three Groups

Source	df	η^2		ε^2		ω^2	
		SS	η^2	SS	η^2	SS	η^2
Parameter Cohen's d	4	322.8	0.06	948.9	0.20	961.3	0.21
Variance/ Heterogeneity	2	46.4	0.01	27.0	0.01	28.7	0.01
Group Size Ratios	3	1.7	0.00	5.3	0.00	5.3	0.00
Population Shape	2	0.3	0.00	0.0	0.00	0.0	0.00
Total N	1	244.1	0.05	254.6	0.05	235.6	0.05
Cohen's d *Variance	8	35.7	0.01	49.1	0.01	51.3	0.01
Cohen's d *Group Prop	12	0.6	0.00	0.9	0.00	0.9	0.00
Cohen's d *Shape	8	0.2	0.00	0.1	0.00	0.1	0.00
Cohen's d *Total N	4	14.1	0.00	0.0	0.00	0.0	0.00
Variance*Group Prop	6	32.3	0.01	3.0	0.00	2.6	0.00
Variance*Shape	4	1.4	0.00	0.8	0.00	0.8	0.00
Variance*Total N	2	17.5	0.00	2.6	0.00	2.4	0.00
Group Prop*Shape	6	0.1	0.00	0.0	0.00	0.0	0.00
Group Prop*Total N	3	0.1	0.00	0.2	0.00	0.2	0.00
Shape*Total N	2	0.1	0.00	0.0	0.00	0.0	0.00
Cohen's d *Variance*Group Prop	24	10.1	0.00	9.1	0.00	9.3	0.00
Cohen's d *Variance*Shape	16	0.7	0.00	0.5	0.00	0.5	0.00
Cohen's d *Variance*Total N	8	6.9	0.00	2.4	0.00	2.3	0.00
Cohen's d * Group Prop*Shape	24	0.1	0.00	0.1	0.00	0.1	0.00
Cohen's d * Group Prop*Total N	12	0.2	0.00	0.1	0.00	0.1	0.00
Cohen's d * Shape *Total N	8	0.1	0.00	0.0	0.00	0.0	0.00
Variance* Group Prop *Shape	12	0.0	0.00	0.1	0.00	0.1	0.00
Variance* Group Prop *Total N	6	7.7	0.00	1.5	0.00	1.4	0.00
Variance* Shape*Total N	4	0.2	0.00	0.1	0.00	0.1	0.00
Group Prop* Shape*Total N	6	0.0	0.00	0.0	0.00	0.0	0.00
Cohen's d * Variance* Group Prop*Shape	48	0.1	0.00	0.1	0.00	0.1	0.00
Cohen's d * Variance* Group Prop*Total N	24	0.0	0.00	0.0	0.00	0.1	0.00
Cohen's d *Variance*Shape*Total N	16	0.1	0.00	0.0	0.00	0.0	0.00
Cohen's d *Group Prop*Shape*Total N	24	0.0	0.00	0.1	0.00	0.1	0.00
Variance* Group Prop*Shape*Total N	12	0.0	0.00	0.0	0.00	0.0	0.00
Cohen's d * Variance* Group Prop*Shape*Total N	48	0.1	0.00	0.1	0.00	0.1	0.00
Error	1799640	4498.0	0.86	3347.0	0.72	3232.9	0.71
Total	1799999	5241.0		4654.0		4536.2	

ANOVA summary Table 21 details that for the $k=3$ condition, Cohen's d was not responsible for as much of the total variance observed in the absolute parameter bias as it

had been in the two group condition. However, in comparison to the parameter bias summary ANOVA table given in Table 18, the amount of variance explained in the absolute parameter bias by Cohen's d for η^2 has gone down by 3% and more than doubled for ϵ^2 and ω^2 to 20% and 21%, respectively. In addition total N as a design condition is accounting for 5% of the total variance per effect size. The remainder of the design conditions and their interactions accounted for less than 5% of the total variance observed across all effect sizes.

For $k = 4$, Cohen's d played an even smaller role in the total amount of variance present in the absolute parameter bias, by accounting for only 1%, 16% and 16% of the total variance across η^2 , ϵ^2 and ω^2 , respectively as shown in Table 22. Also noteworthy in the $k = 4$ condition, the total N design condition showed a comparatively larger percentage of the explained variance present in the absolute parameter bias not only for η^2 but for ϵ^2 and ω^2 as well, accounting for 7%, 7% and 6%, respectively, of the total variance explained.

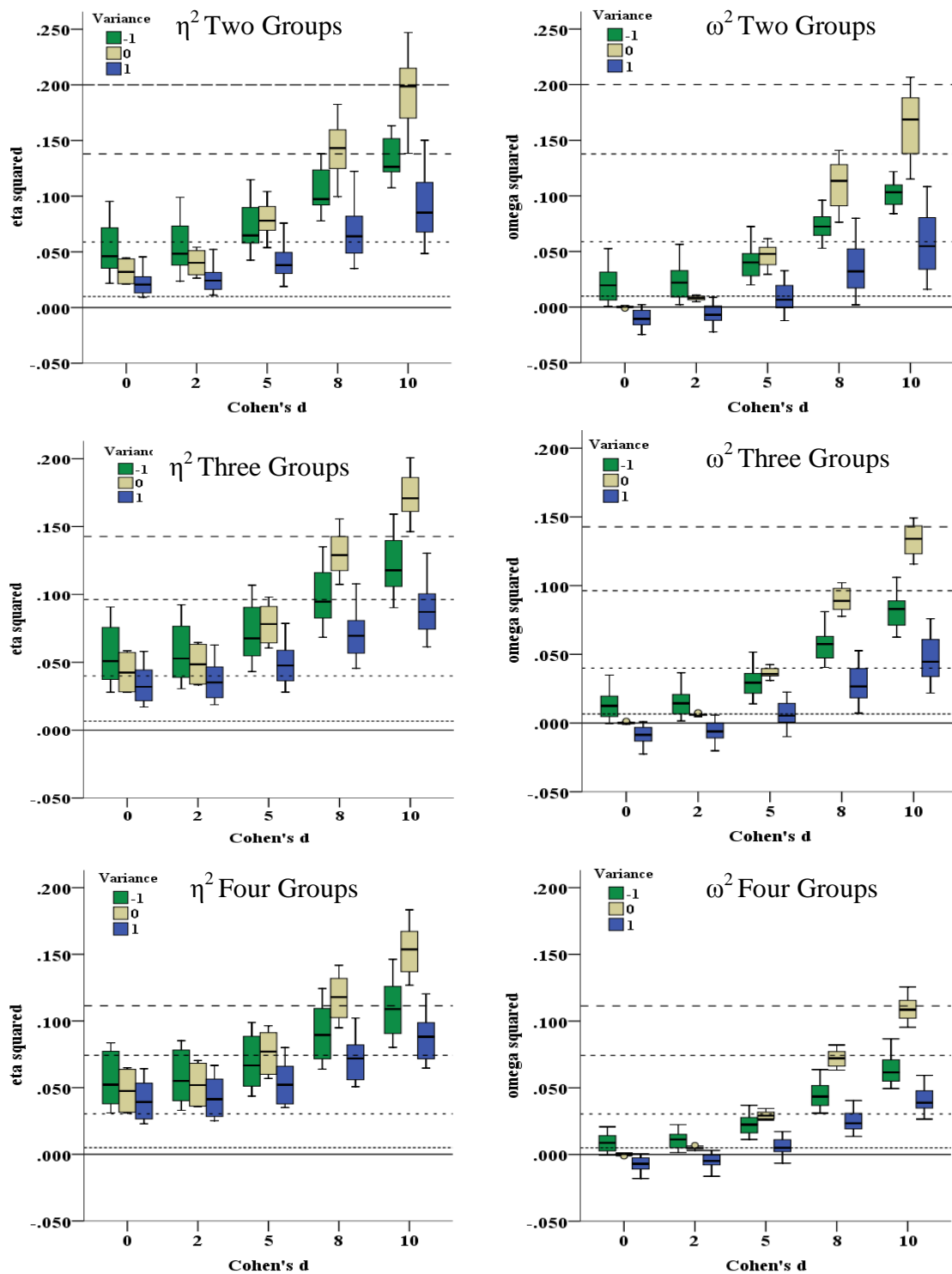
Mean Effect Size Estimates

The negative variance conditions and the positive variance conditions depicted in Figure 6 demonstrate what happens to measures of magnitude of effect when one or more of the groups have more variability than the other groups. For the $k = 2$ condition, Cohen's d values up to 0.2 tend to produce overestimated population eta squared values, and Cohen's d values of 0.8 and 1.0 tend to underestimate population eta squared values under heterogeneity of variance. Similarly for the $k = 3$ condition, Cohen's d values up

Table 22
Estimated Absolute Bias in Three ANOVA Effect Sizes for Four Groups

Source	df	η^2		ε^2		ω^2	
		SS	η^2	SS	η^2	SS	η^2
Parameter Cohen's <i>d</i>	4	56.3	0.01	493.5	0.16	498.4	0.16
Variance/Heterogeneity	2	49.4	0.01	11.9	0.00	12.7	0.00
Group Size Ratios	3	1.8	0.00	0.8	0.00	0.8	0.00
Population Shape	2	0.3	0.00	0.0	0.00	0.0	0.00
Total <i>N</i>	1	303.7	0.07	208.1	0.07	195.9	0.06
Cohen's <i>d</i> *Variance	8	22.4	0.01	21.6	0.01	22.6	0.01
Cohen's <i>d</i> *Group Prop	12	1.2	0.00	0.1	0.00	0.2	0.00
Cohen's <i>d</i> *Shape	8	0.2	0.00	0.1	0.00	0.0	0.00
Cohen's <i>d</i> *Total <i>N</i>	4	17.4	0.00	0.0	0.00	0.0	0.00
Variance*Group Prop	6	19.4	0.00	0.4	0.00	0.3	0.00
Variance*Shape	4	0.9	0.00	0.5	0.00	0.4	0.00
Variance*Total <i>N</i>	2	10.7	0.00	0.8	0.00	0.8	0.00
Group Prop*Shape	6	0.0	0.00	0.0	0.00	0.0	0.00
Group Prop*Total <i>N</i>	3	0.1	0.00	0.2	0.00	0.1	0.00
Shape*Total <i>N</i>	2	0.1	0.00	0.0	0.00	0.0	0.00
Cohen's <i>d</i> *Variance*Group Prop	24	2.8	0.00	3.2	0.00	3.3	0.00
Cohen's <i>d</i> *Variance*Shape	16	0.4	0.00	0.2	0.00	0.2	0.00
Cohen's <i>d</i> *Variance*Total <i>N</i>	8	4.9	0.00	1.3	0.00	1.2	0.00
Cohen's <i>d</i> * Group Prop*Shape	24	0.1	0.00	0.0	0.00	0.0	0.00
Cohen's <i>d</i> * Group Prop*Total <i>N</i>	12	0.2	0.00	0.0	0.00	0.0	0.00
Cohen's <i>d</i> * Shape *Total <i>N</i>	8	0.1	0.00	0.0	0.00	0.0	0.00
Variance* Group Prop *Shape	12	0.0	0.00	0.1	0.00	0.1	0.00
Variance* Group Prop *Total <i>N</i>	6	3.6	0.00	0.4	0.00	0.3	0.00
Variance* Shape*Total <i>N</i>	4	0.1	0.00	0.0	0.00	0.0	0.00
Group Prop* Shape*Total <i>N</i>	6	0.0	0.00	0.0	0.00	0.0	0.00
Cohen's <i>d</i> * Variance* Group Prop*Shape	48	0.1	0.00	0.1	0.00	0.1	0.00
Cohen's <i>d</i> * Variance* Group Prop*Total <i>N</i>	24	0.1	0.00	0.1	0.00	0.1	0.00
Cohen's <i>d</i> *Variance*Shape*Total <i>N</i>	16	0.0	0.00	0.0	0.00	0.0	0.00
Cohen's <i>d</i> *Group Prop*Shape*Total <i>N</i>	24	0.1	0.00	0.0	0.00	0.0	0.00
Variance* Group Prop*Shape*Total <i>N</i>	12	0.0	0.00	0.0	0.00	0.0	0.00
Cohen's <i>d</i> * Variance* Group Prop*Shape*Total <i>N</i>	48	0.1	0.00	0.1	0.00	0.1	0.00
Error	1799640	3651.8	0.88	2397.1	0.76	2329.9	0.76
Total	1799999	4148.2		3140.6		3067.6	

Figure 6. Empirical Estimates of Eta Squared and Omega Squared Under Variance Conditions by Group Size Configuration



to 0.2 tend to produce overestimated population eta squared values, and Cohen's d values of 1.0 tend to underestimate population eta squared values under heterogeneity of variance. For the $k = 4$ condition, Cohen's d values up to 0.5 tend to produce overestimated population eta squared values under heterogeneity of variance.

Also given in Figure 6 are the omega squared values which are supposed to give unbiased estimates of the population eta squared value. The graphs for epsilon squared were practically equivalent to the omega squared graphs, and are thus not presented here. Under conditions of heterogeneity of variance, except for Cohen's $d = 0$ and $d = 0.2$, both the positive and negative variance condition resulted in underestimated parameter eta square values. In the case of Cohen's $d = 0$ and $d = 0.2$, the negative variance pairing overestimated the parameter eta square values, while the positive variance pairing underestimated the parameter eta square.

Precision

Measures of dispersion can be used to “(a) characterize how well location descriptive statistics perform at representing all the data, and (b) to characterize score ‘spreadoutness’ as an important result in its own right” (Thompson, 2006, p. 53). The next set of ANOVA summary tables use standard deviation as an estimate of precision to both complement the preceding results and as a noteworthy statistic in its own right.

While less than half of the total variance was accounted for by the design conditions for parameter bias and absolute parameter bias across any of the group size configurations, this is not the case when standard deviation is the dependent variable. In the $k = 2$ condition, as detailed in Table 23, four design conditions show variance

explained proportions greater than 10%, Cohen's d , total N , variance, and variance by group proportion.

Table 23
Two Group Precision

Source	df	η^2		ω^2		ε^2	
		SS	η^2	SS	η^2	SS	η^2
Paramter Cohen's d	4	.168	.408	.174	.405	.179	.397
Variance/Heterogeneity	2	.079	.191	.081	.190	.084	.187
Group Size Ratios	3	.000	.000	.000	.000	.000	.000
Population Shape	2	.000	.001	.000	.001	.000	.001
Total N	1	.096	.233	.101	.236	.113	.250
Cohen's d *Variance	8	.008	.019	.008	.019	.008	.018
Cohen's d *Group Prop	12	.002	.005	.002	.005	.002	.005
Cohen's d *Shape	8	.000	.001	.000	.001	.000	.001
Cohen's d *Total N	4	.001	.001	.001	.002	.001	.002
Variance*Group Prop	6	.048	.118	.050	.117	.052	.115
Variance*Shape	4	.002	.005	.002	.005	.002	.005
Variance*Total N	2	.003	.007	.003	.008	.004	.008
Group Prop*Shape	6	.000	.000	.000	.000	.000	.000
Group Prop*Total N	3	.000	.000	.000	.000	.000	.000
Shape*Total N	2	.000	.000	.000	.000	.000	.000
Error	292	.004	.011	.005	.011	.005	.011
Total	359	.411		.428		.451	

Estimated Precision

In the $k = 3$ condition, as detailed in Table 24, the same four design conditions show variance explained proportions greater than 5% as were evidenced in Table 23. However, unlike Table 23, total N has almost as much, if not more, predictive power than Cohen's d in determining the precision of mean effect size estimates across η^2 , ε^2 and ω^2 for the $k = 3$ condition.

Table 24
Three Group Precision

Source	df	η^2		ω^2		ε^2	
		SS	η^2	SS	η^2	SS	η^2
Parameter Cohen's d	4	.081	.398	.086	.387	.088	.379
Variance/Heterogeneity	2	.023	.116	.025	.113	.026	.110
Group Size Ratios	3	.000	.000	.000	.000	.000	.000
Population Shape	2	.000	.001	.000	.001	.000	.001
Total N	1	.076	.374	.087	.389	.094	.402
Cohen's d *Variance	8	.004	.022	.005	.021	.005	.021
Cohen's d *Group Prop	12	.000	.001	.000	.001	.000	.001
Cohen's d *Shape	8	.000	.001	.000	.001	.000	.001
Cohen's d *Total N	4	.000	.001	.000	.001	.000	.001
Variance*Group Prop	6	.014	.070	.015	.068	.016	.067
Variance*Shape	4	.001	.004	.001	.004	.001	.004
Variance*Total N	2	.001	.005	.001	.006	.001	.006
Group Prop*Shape	6	.000	.000	.000	.000	.000	.000
Group Prop*Total N	3	.000	.000	.000	.000	.000	.000
Shape*Total N	2	.000	.000	.000	.000	.000	.000
Error	292	.002	.008	.002	.008	.002	.008
Total	359	.203		.223		.233	

Note. All higher order interactions were excluded because the percent of variance explained for each was less than 0.01.

In the $k = 4$ condition, as detailed in Table 25, Cohen's d , variance and total N design conditions show variance explained proportions greater than 5% . Now total N is more indicative of the amount of precision than Cohen's d across η^2 , ε^2 and ω^2 for the $k = 4$ condition.

Table 25
Four Group Precision

Source	df	η^2		ω^2		ε^2	
		SS	η^2	SS	η^2	SS	η^2
Paramter Cohen's d	4	.046	.379	.050	.362	.051	.354
Variance/Heterogeneity	2	.008	.063	.008	.060	.008	.059
Group Size Ratios	3	.000	.001	.000	.001	.000	.001
Population Shape	2	.000	.000	.000	.000	.000	.000
Total N	1	.060	.490	.071	.511	.075	.521
Cohen's d *Variance	8	.002	.020	.003	.019	.003	.018
Cohen's d *Group Prop	12	.000	.001	.000	.001	.000	.001
Cohen's d *Shape	8	.000	.001	.000	.001	.000	.001
Cohen's d *Total N	4	.000	.001	.000	.001	.000	.001
Variance*Group Prop	6	.004	.034	.004	.032	.005	.032
Variance*Shape	4	.000	.003	.000	.003	.000	.003
Variance*Total N	2	.000	.003	.000	.003	.000	.003
Group Prop*Shape	6	.000	.000	.000	.000	.000	.000
Group Prop*Total N	3	.000	.000	.000	.000	.000	.000
Shape*Total N	2	.000	.000	.000	.000	.000	.000
Error	292	.001	.005	.001	.005	.001	.005
Total	359	.122		.138		.144	

Note. All higher order interactions were excluded because the percent of variance explained for each was less than 0.01.

Discussion

The present findings provide evidence that Cohen's d impacts the parameter bias estimates across all group size configurations, although less so as the number of groups being compared increases. The impact of variance on parameter biases is less than that of Cohen's d but remains relatively consistent across the group size conditions. Similarly the parameter biases across η^2 , ε^2 and ω^2 appears relatively consistent if not somewhat improved for ε^2 and ω^2 within each group size condition. Thus, in light of the parameter

bias only, there appears to be no real detriment or advantage to using correction formulas for η^2 in terms of parameter bias.

Absolute parameter biases however, paint a different picture. In each group size condition, η^2 accounts for a smaller portion of the total absolute parameter bias than ϵ^2 and ω^2 . Thus taken together, the overestimated and underestimated estimates of parameter bias are greater when using correction formulas ϵ^2 and ω^2 than if no correction formula was used! Variance no longer offers explanatory power for this absolute bias. Instead, total N appears to offer more of an explanation for the variability in absolute parameter bias.

Parameter bias and absolute parameter bias estimates are a useful yet partial depiction of the story. A more complete account further examines the dispersion about the parameter estimates, the standard deviation, to provide an index of precision. Simulation results reveal that a large majority of the variability observed in the precision of the estimates is accountable by the design conditions. Specifically, there is empirical support that Cohen's d , variance, total N and variance by group proportion play a large role in understanding the variability in the precision of η^2 , ϵ^2 and ω^2 estimates. In fact, as the number of groups increases, total N plays a larger role than Cohen's d in explaining the variability that is present in the precision of the estimates of η^2 , ϵ^2 and ω^2 .

So What?

While it is well known that η^2 is a positively biased measure of effect, researchers may be unaware that the use correction formulas may actually result in *poorer* estimates of population effect size. The results of the present study make clear that depending on

the total N and Cohen's d value, ϵ^2 and ω^2 actually produce *more* absolute parameter bias than η^2 . However, it is not the intent of the present paper to advocate for the discontinued use of either correction formulas, but rather that each researcher carefully considers the present finding before making the decision to correct or not to correct.

Just as researchers recognize assumptions violations impact the validity of ANOVA results for measures of statistical significance; researchers need to recognize that assumption violations similarly impact measures of practical significance. Thus, design conditions of variance and variance by group proportion impact the estimated precision of estimates of η^2 , ϵ^2 and ω^2 . Moreover, Cohen's d and total N account for the much of the variability that is present in effect size estimates. Total N is especially indicative of the precision in the estimates as the number of groups increases from $k = 2$ to $k = 4$.

Statistical analysis is a minds-on endeavor. As Thompson (2006) noted, “good social science research is primarily about *thinking*, about *reflection*, and about *judgment* [emphasis added]” (v). Before making analytical tool selections, researchers must understand the distributional characteristics of their data at hand. Only after understanding one's own data, can one understand what analytical tool will most accurately and precisely tell the story of the data at hand. Each data has its own unique characteristics. In the end, the choice that is both well supported from previous research efforts and has taken account the idiosyncratic nature of the data is arguably a better choice, than any given rule of thumb.

SUMMARY AND CONCLUSIONS

Today's researchers have a whole array of analytical tools to investigate their data. As types of statistical techniques present in the literature continue to grow, unifying concepts such as the GLM allow for vital connections among the seemingly disconnected array of statistical techniques. As described in previous sections, parametric techniques subsumed under the GLM share the unifying concepts that all statistical analyses are (a) correlational (b) apply weights to measured variables to obtain latent variables and (c) yield effect sizes analogous to r^2 (Thompson, 2006).

While the numbers of analytical tools has increased, the number of research articles published has increased at an incomprehensible rate. Reviews of the literature make an important contribution when they cogently synthesize previous research findings to paint a picture that was not visible by examining individual studies in isolation. The first study examined historical trends in data analytical technique choices in educational and psychological literature. Only through continued reviews of the literature and periodic synthesis of these reviews can we begin to understand the direction and possible motivating forces behind changes in analytical practices.

Study One

Findings from the first study underscore shifts in both the education and psychology, though not necessarily in the same direction. Cohen's (1968) seminal piece raised awareness of the GLM in a time when ANOVA practices were widely popular. A decade later Knapp further raised awareness in the GLM by describing canonical

correlation analysis (CCA) as the more general case of the GLM. Next in 1981, Bagozzi, Fornell and Larcker explained that CCA was a special case of SEM. Figure 1 details not only the downward trends in both t test and ANOVA/ANCOVA practices in education but also the upward trends in regression practices. Perhaps, recognition of GLM principles are finally becoming incorporated into research practices as researchers understand that ANOVA is not inherently superior by design to regression.

Current trends in psychology, as detailed in Table 6, point to correlation as the most popular technique. On the other hand ANOVA is currently the most popular technique in educational literature. Teaching practices, reviewing processes, and reporting recommendations need to consider where current practices are now as decisions are made regarding future data analytic directions.

A practice that finally appears to be gaining momentum in the literature is the practice of using and reporting effect size measures. Understanding the connections among the GLM should direct researchers to the reality that that variance accounted for effect sizes such as Pearson r^2 squared and η^2 are biased estimators of population effect sizes. Study two and three addressed correction of these biases and evaluated the impact of assumption violations on computed estimates.

Study Two

The second study contributes to the literature by offering empirical evidence for choosing the most appropriate correction formula for the Pearson r^2 . Study two is particularly relevant not only because correlational techniques are currently the most popular technique in psychology but also because of current reform efforts where

researchers are acknowledging the need to report measures of magnitude of effect. Findings from the second study report that the most popular correction formula commonly used in correcting R^2 , the Ezekiel formula, is not necessarily the best correction formula to use. Instead, the Pratt formula may be a better choice across design features investigated. Yin and Fan (2001) similarly considered the Pratt formula superior to the Ezekiel in their evaluation of multiple R^2 formulas. The Olkin-Pratt Extended performed very well also. Cattin (1980) recommended the use of the Olkin-Pratt Extended over the Ezekiel for correction of r^2 .

Study Three

In the third and final study, effect sizes for ANOVA, the most popular technique in education, were considered under conditions when statistical assumptions were not met. Not only is this study important in terms of the current reform movement's increased concern for effect size estimates, but perhaps more importantly in terms of understanding that assumption violations not only impact estimates of statistical significance, such as Type I error rates and power, but can also distort measures of magnitude of effect. Thus, an investigation of the behavior of estimates of practical significance conveys important and timely information for a large number of studies. Findings, as detailed in Figure 6, from the third study demonstrate that η^2 for the $k = 2$ condition at Cohen's d values up to 0.2 tends to produce overestimated population η^2 values, and Cohen's d values of 0.8 and 1.0 tend to underestimate population η^2 under heterogeneity of variance. For the $k = 3$ condition, Cohen's d values up to 0.2 tend to produce overestimated population η^2 values, and Cohen's d values of 1.0 tend to

underestimate population η^2 under heterogeneity of variance. For the $k = 4$ condition, Cohen's d values up to 0.5 tend to produce overestimated population η^2 under heterogeneity of variance. When the homogeneity of variance assumption was met, only at Cohen's d values of 0.8 and 1.0 for the $k=2$ condition was the population η^2 not consistently overestimated.

Computed ε^2 and ω^2 where Cohen's $d=0$ and $d=0.2$, the negative variance pairing overestimated the parameter η^2 values, while the positive variance pairing underestimated the parameter η^2 under conditions of heterogeneity of variance. However, in the cases of Cohen's $d \geq 0.5$, both the positive and negative variance conditions resulted in underestimated parameter η^2 values for the computed ε^2 and ω^2 . When the homogeneity of variance assumption was met, at $k=2$ and $k=3$ condition, computed ε^2 and ω^2 at Cohen's d values greater than 0.2 consistently produced underestimated population η^2 values. On the other hand, under homogeneity of variance, computed ε^2 and ω^2 at the $k=4$ condition consistently produced accurate estimates of the population η^2 values across all the Cohen's d values investigated.

In closing, a general recommendation from all three studies is that researchers should make an effort to be transparent about their particular data characteristics *and* statistical technique choices. As a research community we need to critically examine these analytic choices in the literature considering best practices recommendations. As researchers, we must hold ourselves accountable for defending our analytical decisions. Although there are many equations in statistics, there is not a single prescriptive equation that definitively tells researchers which analytical decision is best. Instead, each

researcher serves as the content expert of their own study and must provide evidence that the best possible analytical choices were made to best describe the data at hand.

REFERENCES

- Aiken, L. S., West, S. G., & Millsap, R. E. (2008). Doctoral training in statistics, measurement and methodology in psychology. *American Psychologist*, 63, 32-50.
- Aiken, L. S., West, S. G., Sechrest, L., & Reno, R. R. (1990). Graduate training in statistics, methodology, and measurement in psychology. *American Psychologist*, 45, 721-734.
- American Educational Research Association (AERA). (2006). Standards for reporting on empirical social science research in AERA publications. *Educational Researcher*, 35(6), 33-40.
- American Psychological Association (APA). (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: American Psychological Association.
- American Psychological Association (APA). (2006). Evidence-based practice in psychology. *American Psychologist*, 61, 271-285.
- American Psychological Association (APA). (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.
- Bagozzi, R. P., Fornell, C., & Larcker, D. F. (1981). Canonical correlation analysis as a special case of a structural relations model. *Multivariate Behavioral Research*, 16, 437-454.

- Bangert, A. W., & Baumberger, J. P. (2005). Research and statistical techniques used in the *Journal of Counseling & Development*: 1990-2001. *Journal of Counseling & Development*, 83, 480-487.
- Bliss, S. L., Skinner, C. H., Hautau, B., & Carroll, E. E. (2008). Articles published in four school psychology journals from 2000 to 2005: An analysis of experimental/intervention research. *Psychology in the Schools*, 45, 483-498.
- Bowen, B. E., Rollins, T. J., Baggett, C. D., & Miller, J. P. (1990). Statistical procedures employed in the journal of agricultural education. *Journal of Agricultural Education*, 31.
- Capraro, M. M., & Capraro, R. M. (2003). Exploring the APA fifth edition Publication Manual's impact on the analytic preferences of journal editorial board members. *Educational and Psychological Measurement*, 63(4), 554.
- Capraro, M. M., Capraro, R. M., & Henson, R. K. (2001). Measurement error of scores on the mathematics anxiety rating scale across studies. *Educational and Psychological Measurement*, 61, 373-386.
- Capraro, R. M., & Thompson, B. (2008). The educational researcher defined: What will future researchers be trained to do? *The Journal of Educational Research*, 10, 247-253.
- Carroll, R. M., & Nordholm, L. A. (1975). Sampling characteristics of Kelley's ϵ^2 and Hay's ω^2 . *Educational and Psychological Measurement*, 35, 541-554.
- Cattin, P. (1980). Estimation of the predictive power of a regression model. *Journal of Applied Psychology*, 65, 407-414.

- Chow, S. L. (1996). *Statistical significance: Rationale, validity and utility*. London: Sage.
- Claudy, J. G. (1978). Multiple regression and validity estimation in one sample. *Applied Psychological Measurement*, 2, 595-607.
- Cleveland, W. S., & Devlin, S. J. (1988). Locally-weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83, 596-610.
- Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin*, 70, 426-443.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
- Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, 45, 1304-1312.
- David, H. A. (1995). First (?) occurrence of common terms in mathematical statistics. *The American Statistician*, 49, 121-133.
- Donaldson, T. S. (1968). Robustness of the F -test to errors of both kinds and the correlation between the numerator and denominator of the F -ratio. *Journal of the American Statistical Association*, 63, 660-676.
- Edington, E. S. (1964). A tabulation of inferential statistics used in psychology journals. *American Psychologist*, 19, 202-203.
- Edington, E. S. (1974). A new tabulation of statistical procedures used in APA journals. *American Psychologist*, 25-26.

- Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7, 1-26.
- Elmore, P. B., & Woehlke, P. L. (1988). Statistical methods employed in "American Educational Research Journal", "Educational Researcher", and "Review of Educational Research" from 1978-1987. *Educational Researcher*, 17(9), 19-20.
- Elmore, P. B., & Woehlke, P. L. (1996, April). *Research methods employed in American Educational Research Journal, Educational Researcher, and Review of Educational Research from 1978-1995*. Paper presented at the Annual Meeting of the American Educational Research Association, New York.
- Elmore, P. B., & Woehlke, P. L. (1998, April). *Twenty years of research methods employed in "American Educational Research Journal," "Educational Researcher," and "Review of Educational Research"*. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.
- Emmons, N. J., Stallings, W. M., & Layne, B. H. (1990, April). *Statistical methods used in "American Educational Research Journal", "Journal of Educational Psychology," and "Sociology of Education" from 1972 through 1987*. Paper presented at the Annual Meeting of the American Educational Research Association, Boston, MA.
- Erceg-Hurn, D. M., & Mirosevich, V. M. (2008). Modern robust statistical methods. *American Psychologist*, 63, 591-601.

- Erwin, J. (2007). Statistical reform in psychology. *Observer*, 20(4), 1. Retrieved from <http://www.psychologicalscience.org/observer/getArticle.cfm?id=2149>
- Ezekiel, M. (1929). The application of the theory of error to multiple and curvilinear correlation. *Journal of the American Statistical Association*, 24, 99-104.
- Ezekiel, M. (1930). *Methods of correlational analysis*. New York: Wiley.
- Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4, 272-299.
- Fan, X., Felsovalyi, A., Sivo, S. A., & Keenan, S. C. (2001). *SAS for Monte Carlo studies: A guide for quantitative researchers*. Cary, NC: SAS Institute.
- Fan, X., & Thompson, B. (2001). Confidence intervals for effect sizes. *Educational and Psychological Measurement*, 61, 517-531.
- Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral and biomedical sciences. *Behavior Research Methods*, 39, 175-191.
- Fidler, F. (2002). The fifth edition of the APA Publication Manual: Why its statistics recommendations are so controversial. *Educational and Psychological Measurement*, 62, 749-770.
- Finch, S., Thomason, N., & Cumming, G. (2002). Past and future American Psychological Association guidelines for statistical practice. *Theory & Psychology*, 12, 825-853.

- Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh, Scotland: Oliver and Boyd.
- Fisher, R. A., & Mackenzie, W. A. (1923). The manurial response of different potato varieties. *Journal of Agricultural Science*, 13, 311-320.
- Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika*, 43, 521-532.
- Friedrich, J., & Buday, E. (2000). Statistical training in psychology: A national survey and commentary on undergraduate programs. *Teaching of Psychology*, 27, 248-257.
- Gamst, G., Meyers, L. S., & Guarino, A. J. (2008). *Analysis of variance designs: A conceptual and computational approach with SPSS and SAS*. New York: Cambridge University Press.
- Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumption underlying the fixed effects analysis of variance and covariance. *Review of Educational Research*, 42, 237- 288.
- Goodwin, L. D., & Goodwin, W. L. (1985a). An analysis of statistical techniques used in the *Journal of Educational Psychology*, 1979-1983. *Educational Psychologist*, 20, 13-21.
- Goodwin, L. D., & Goodwin, W. L. (1985b). Statistical techniques in *AERJ* articles, 1979-1983: The preparation of graduate students to read the educational research literature. *Educational Researcher*, 14(2), 5-11.

- Grissom, R. J., & Kim, J. J. (2005) *Effect sizes for research: A broad practical approach*. New York: Psychology Press.
- Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (1997). *What if there were no significance tests*. Mahwah, NJ: Lawrence Erlbaum.
- Harwell, M. R., Rubinstein, E. N., Hayes, W. S., & Olds, C. C. (1992). Summarizing Monte Carlo results in methodological research: The one- and two- factor fixed effects ANOVA cases. *Journal of Educational Statistics*, 17, 315-339.
- Hays, W. L. (1981). *Statistics* (3rd ed.). New York: Holt, Rinehart & Winston.
- Hlebowitsh, P. S. (2005). *Designing the School Curriculum*. Boston: Pearson Education.
- Horton, N. J., & Switzer, S. S. (2005). Statistical methods in the Journal. *New England Journal of Medicine*, 353, 1977-1979.
- Hubbard, R., Bayarri, M. J., Berk, K. N., & Carlton, M. A. (2003). Confusion over measures of evidence (p 's) versus errors (α 's) in classical statistical testing. *The American Statistician*, 57, 171-182.
- Hubbard, R., & Ryan, P. A. (2000). The historical growth of statistical significance testing in psychology - and its future prospects. *Educational and Psychological Measurement*, 60, 661-681.
- Huberty, C. J., & Mourad, S. A. (1980). Estimation in multiple correlation/prediction. *Educational and Psychological Measurement*, 40, 101-112.
- Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Thousand Oaks, CA: Sage Publications.

- Journal of Experimental Education*. (1993). Special Issue - "The role of statistical significance testing in contemporary analytical practice: Alternatives with commends from journal editors". Washington, DC: Heldref Publications.
- Kaufman, A. S. (1998). Introduction to the special issue on statistical significance testing *Research in the Schools*, 5, 1.
- Keiffer, K. M., Reese, R. J., & Thompson, B. (2001). Statistical techniques employed in AERJ and JCP articles from 1988 to 1997: A methodological review. *Journal of Experimental Education*, 69, 280-309.
- Kelley, T. L. (1935). An unbiased correlation ratio measure. *Proceedings of the National Academy of Sciences*, 21, 554-559.
- Keselman, H. J. (1975). A Monte Carlo investigation of three estimates of treatment magnitude: Epsilon squared, eta squared, and omega squared. *Canadian Psychological Review*, 16, 44-48.
- Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B. et al. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research*, 68, 350-386.
- Kieffer, K. M., Reese, R. J., & Thompson, B. (2001). Statistical techniques employed in "AERJ" and "JCP" articles from 1988 to 1997: A methodological review. *Journal of Experimental Education*, 69, 280-309.
- Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, 56, 746-759.

- Kline, R. B. (2009). *Becoming a behavioral science researcher*. New York: Guilford.
- Knapp, T. R. (1978). Canonical correlation analysis: A general parametric significance-testing system. *Psychological Bulletin*, 85, 410-416.
- Kromney, J. D., & Hines, C. V. (1996). Estimating the coefficient of cross-validity in multiple regression: A comparison of analytical and empirical methods. *The Journal of Experimental Education*, 64, 240-266.
- Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47, 583-621.
- Lix, L. M., Keselman, J. C., & Keselman, H. J. (1996). Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance "F" test. *Review of Educational Research*, 66, 579-619.
- Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, 105, 156-164.
- Mohr, L. B. (1990). *Understanding significance testing*. Newbury Park, CA: SAGE.
- Morrison, D. E., & Henkel, R. E. (1973). *The significant test controversy: A reader*. Chicago: Aldine Publishing Company.
- Natesan, P., & Thompson, B. (2007). Extending improvement-over-chance I-index effect size simulation studies to cover some small-sample cases. *Educational and Psychological Measurement*, 67, 59-72.
- Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika*, 29A, Part I: 175-240; part II 263-294.

Olkin, I., & Pratt, J. W. (1958). Unbiased estimation of certain correlation coefficients.

The Annals of Mathematical Statistics, 29, 201-211.

Pearson, E. S. (1931). The analysis of variance in cases of non-normal variation.

Biometrika, 23, 114-133.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space.

Philosophical Magazine, 6, 559-572.

Raju, N. S., Bilgic, R., Edwards, J. E., & Fleer, P. F. (1999). Accuracy of population

validity and cross-validity estimation: An empirical comparison of formula-

based, traditional empirical, and equal weights procedure. *Applied Psychological*

Measurement, 23, 99-115.

Robey, R. R., & Barcikowski, R. S. (1992). Type I error and the number of iterations in

Monte Carlo studies of robustness. *British Journal of Mathematical and*

Statistical Psychology, 45, 283-288.

Robinson, D. H., Levin, J. R., Thomas, G. D., Pituch, K. A., & Vaughn, S. (2007). The

incidence of "causal" statements in teaching-and-learning research journals.

American Educational Research Journal, 44, 400-413.

Schinka, J. A., LaLone, L., & Broeckel, J. A. (1997). Statistical methods in personality

assessment research. *Journal of Personality Assessment*, 68, 487-496.

Shavelson, R. J., & Towne, L. (Eds.). (2002). *Scientific research in education*.

Washington, DC: National Academy Press.

Sigel, S., & Catellan, N. J. (1988). *Nonparametric statistics for the behavioral sciences*.

New York: McGraw-Hill.

- Snyder, P., & Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. *Journal of Experimental Education*, 61, 334-349.
- Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology*, 15, 201-293.
- Strasak, A. M., Zaman, Q., Marinell, G., Pfeiffer, K. P., & Ulmer, H. (2007). The use of statistics in medical research: A comparison of The New England Journal of Medicine and Nature Medicine. *American Statistician*, 61, 47-55. doi: 10.1198/000313007x170242
- Stuart, A., Ord, J. K., & Arnold. (Eds.). (1994). *Kendall's advanced theory of statistics: Distribution theory*. New York: Halsted Press.
- Tabachnick, B. G., & Fidell, L. S. (2007). *Experimental designs using ANOVA*. Belmont, CA: Thomson Higher Education.
- Thompson, B. (1994). Guidelines for authors. *Educational and Psychological Measurement*, 54, 837-847.
- Thompson, B. (1997). Statistical significance testing practices in The Journal of Experimental Education. *Journal of Experimental Education*, 66, 75-83.
- Thompson, B. (1998, April). *Five methodology errors in educational research: The pantheon of statistical significance and other faux pas*. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.
- Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher*, 31(3), 25-32.

- Thompson, B. (2006). *Foundations of behavioral statistics: An insight-based approach*. New York: Guilford.
- Thompson, B. (2008). Bruce Thompson's Home Page. *Hyperlinks* Retrieved June 1, 2009, from <http://www.coe.tamu.edu/~bthompson/>
- Tremblay, P. F., & Gardner, R. C. (1996). On the growth of structural equation modeling in psychological journals. *Structural Equation Modeling: A Multidisciplinary Journal*, 3, 93-104.
- Tukey, J. W. (1960). A survey of sampling from contaminated distributions. In S. G. Olkin, W. Hoeffding, W. Madow & H. Mann (Eds.), *Contributions to probability and statistics: Essays in honor of Harold Hotelling*. Stanford, CA: Stanford University Press.
- Vacha-Haase, T., Kogan, L. R., & Thompson, B. (2000). Sample compositions and variabilities in published studies versus those in test manual: Validity of score reliability inductions. *Educational and Psychological Measurement*, 60, 509-522.
- Vacha-Haase, T., Nilsson, J., Reetz, D., Lance, T., & Thompson, B. (2000). Reporting practices and APA editorial policies regarding statistical significance and effect size. *Theory and Psychology*, 10, 413.
- Vale, C. D., & Maurelli, V. A. (1983). Simulating multivariate nonnormal distributions. *Psychometrika*, 48, 465-471.
- Varnell, S. P., Murray, D. M., Janega, J. B., & Blitstein, J. L. (2004). Design and analysis of group-randomized trials: A review of recent practices. *American Journal of Public Health*, 94, 393-399.

- Walberg, H., Vukosavich, P., & Tsai, S. (1981). Scope and structure of the journal literature in educational research. *Educational Researcher*, 10(8), 11-13.
- Wang, Z., & Thompson, B. (2007). Is the Pearson r^2 biased, and if so, what is the best correction formula? *Journal of Experimental Education*, 75, 109-125.
- Wherry, R. J. (1931). A new formula for predicting the shrinkage of the coefficient of multiple correlation. *The Annals of Mathematical Statistics*, 2, 440-457.
- Whitehurst, G. J. (2002). Archived evidence-based education (EBE) Retrieved December 20, 2007, from <http://www.ed.gov/offices/OERI/presentations/evidencebase.html>
- Wilcox, R. R. (1987). New designs in analysis of variance. *Annual Review of Psychology*, 38, 29-60.
- Wilcox, R. R. (1993). Robustness in ANOVA. In L. K. Edwards (Ed.), *Applied analysis of variance in behavioral science* (pp. 345-374). New York: Marcel Dekker.
- Wilcox, R. R. (1995). ANOVA: A paradigm for low power and misleading measures of effect size. *Review of Educational Research*, 65, 51-77.
- Wilcox, R. R. (2006). Graphical methods for assessing effect size. *Journal of Experimental Education*, 74, 353-367.
- Wilcox, R. R., Charlin, V., & Thompson, K. L. (1986). New Monte Carlo results on the robustness of the ANOVA F, W, F* statistics. *Communications in Statistics: Simulation and Computation*, 15, 933-944.
- Wilcox, R. R., & Keselman, H. J. (2003). Modern robust data analysis methods: Measures of central tendency. *Psychological Methods*, 8, 254-274.

- Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals guidelines and explanations. *American Psychologist*, 54, 594-604.
- Willson, V. (1980). Research techniques in "AERJ" articles: 1969 to 1978. *Educational Researcher*, 9(6), 5-10.
- Yin, P., & Fan, X. (2001). Estimating R^2 shrinkage in multiple regression: A comparison of different analytical methods. *Journal of Experimental Education*, 69, 203-224.
- Zientek, L. R., Capraro, M. M., & Capraro, R. M. (2008). Reporting practices in quantitative teacher education research: One look at the evidence cited in the AERA panel report. *Educational Researcher*, 37(4), 208-216.
- Zimmerman, D. W., Zumbo, B. D., & Williams, R. H. (2003). Bias in estimation and hypothesis testing of correlation. *Psicologica*, 24, 133-158.

VITA

Name: Susana Troncoso Skidmore

Address: Department of Psychology and Anthropology,
College of Social and Behavioral Sciences
1201 W University Drive
Edinburg, TX 78539

Email Address: stskidmore@gmail.com

Education: B.A., Biology, Texas A&M University, 1991
M.Ed., Science Education, Texas A&M University, 2007
Ph.D., Educational Psychology, Texas A&M University, 2009